Looking at Radiology Report Generation through a Causal Lens: A Survey

Anonymous ACL submission

Abstract

Automatic radiology report generation (RRG) has emerged as a promising approach to reduce clinicians' workload, yet existing systems are vulnerable to biases induced by spurious correlations across data, models, and evaluation pipelines. Such biases raise serious fairness concerns and may adversely affect patient care, making their mitigation critical in clinical settings. Leveraging causal inference to identify true cause-effect relationships can mitigate many biases and yield fair, reliable systems with clinically meaningful outputs. Existing surveys on RRG primarily emphasize deep learning approaches while overlooking the critical role of causality. This survey addresses this gap by analyzing bias across the RRG pipeline, formalizing RRG as a causal modeling problem, and reviewing representative causal techniques from the literature. Based on the level of intervention, we organize existing mitigation strategies into a three-tier taxonomy. We further examine commonly used public medical imaging datasets and evaluation metrics through a causal lens, revealing their biases and limitations in capturing causal alignment and clinical fidelity. To address these limitations, we advocate broader demographic coverage and causal-aware evaluation metrics to improve fairness and reliability, and identify important directions for future work.

1 Introduction

Recent advances in deep learning and natural language generation (NLG) enable the automatic translation of medical images into diagnostic text, a task known as *Automated Radiology Report Generation (RRG)*[1] (Artsi et al., 2025). Machine-generated reports, reviewed by radiologists, can accelerate clinical workflows (Liu et al., 2023) amid a global shortage of radiologists (Afshari Mirak et al., 2025; Do et al., 2023; Rimmer, 2017; Arora, 2014). Current RRG methods use deep learning to encode images and large language models (LLMs) to generate text (Wang et al., 2023). Some approaches further integrate knowledge graphs to improve RRG performance (Kale et al., 2022, 2023a,b; Liu et al., 2021). However, performance gaps remain, and owing to the sensitivity of this field, errors can be critical, making fairness and reliability important. Despite technological progress, these systems are vulnerable to biases stemming from data imbalances, cognitive biases in human annotations, and limitations of large language models (LLMs). Such biases can lead to incorrect depiction of health condition in radiology reports thus propagating and amplifying health disparities (e.g., medical devices (e.g., tubes, lines), underrepresented demographic groups, reduced diagnostic accuracy for women), undermining trust and clinical utility. Therefore, mitigating these biases is essential for recovering correct cause-effect relationships in diagnostic reports, motivating the use of causal inference. Further discussion of causal challenges in the medical domain is provided in Appendices J and C.

Causal inference reveals underlying mechanisms, distinguishing genuine cause–effect relationships from mere correlations in observational data (Pearl, 1995). In this survey, we examine RRG from the perspective of causal inference, a direction that is increasingly essential for producing reliable diagnostic reports. A systematic analysis of prior RRG surveys reveals a critical research gap in the integration of causal inference (refer to Appendix B), necessitating this work bridging causal reasoning and automated radiology report generation. Our contributions are:

1. A systematic review of the RRG pipeline, identifying and cataloging potential sources of bias that contribute to report inaccuracy. This analysis forms a critical foundation for under-

[1] In this work, RRG denotes automated radiology report generation.

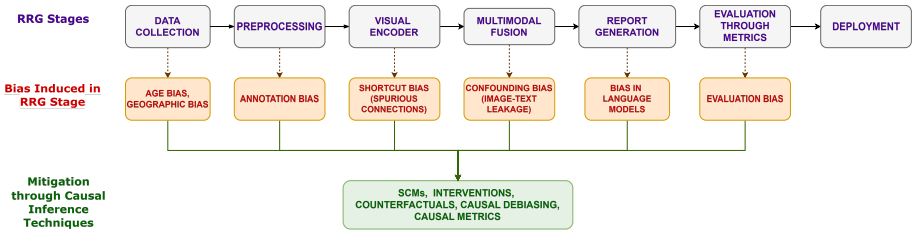

Figure 1: RRG Pipeline with bias induced in each step.

standing RRG bias and guides the design of future mitigation efforts *(Refer §2)*.

2. Modeling RRG as a causal inference problem to mitigate systemic biases, using structural causal models (SCMs) to identify confounding variables and review the application of counterfactual augmentation and causal debiasing. This will help researchers in RRG to apply causal frameworks helping in report accuracy (*Refer §3*).

3. A clear categorization of mitigation approaches into a three-tiered taxonomy: data-level, model-level, and evaluation-level interventions. This systematic categorization offers researchers and practitioners a systematic methodology for selecting and applying optimal interventions to improve RRG fairness and accuracy (*Refer §4*).

4. Examining biases in public medical imaging datasets and evaluation metrics through a causal lens, identifying that these metrics fail to capture causal alignment and clinical fidelity. We argue that broader demographic coverage in datasets (e.g., age and diversity) and causal-aware metrics can help in making RRG more fair and reliable (*Refer §5*).

Figure 3 in Appendix A presents taxonomy used in this survey. It also conveys the outline of this paper.

2 Origin of Bias in RRG

Figure 1 illustrates the complete RRG pipeline with the bias induced at each stage with the solution as causal inference techniques. Bias in RRG originates from various factors but can be broadly classified according to its pipeline into 3 factors: humans, medical imaging and LLMs. These cumulative biases risk perpetuating disparities in diagnosis and treatment recommendations, particularly affecting marginalized populations (Tejani et al., 2024).

2.1 Bias in Radiology Practice and Reporting

The models are trained on data annotated by radiologists who, like all humans, are susceptible to cognitive biases that may influence their interpretation and reporting of imaging studies. The key cognitive biases include the following.

i **Alliterative Bias:** Radiologists rely heavily on previous reports when interpreting follow-up studies, which could perpetuate previous errors or assumptions. This bias can lead to diagnostic inertia, where new findings are overlooked (Busby et al., 2018; Zhang et al., 2023). The study by (Murphy et al., 2024) has shown that it accounts for about 6% of diagnostic errors in radiology.

ii **Framing Bias:** Radiologists' diagnostic conclusions can be influenced by how clinical information is presented. Limited or misleading clinical histories (for example, brief or incomplete indications) can skew interpretation, sometimes causing errors if the whole context is unknown (Busby et al., 2018; Lee et al., 2013).

iii **Availability Bias:** Diagnoses that are more easily recalled or recently encountered tend to be overestimated, e.g., a radiologist who recently missed lung cancer may overcall nodules, causing false positives (Busby et al., 2018; Itri and Patel, 2018).

iv **Anchoring bias:** The undue influence of an initial diagnostic impression on subsequent decision-making despite new contradictory information, is typically viewed as a source of diagnostic error in radiology (Itri and Patel, 2018; Yoon et al., 2024).

v **Hindsight bias:** The tendency to overestimate the predictability of an event after knowing the outcome can affect learning from errors and diagnostic decision making (Itri and Patel, 2018; Onder et al., 2021).

vi **Blind Spot bias:** Radiologists may be aware of common errors but general or misinterpret findings due to increased vigilance, which can also lead to errors (Yoon et al., 2024; Onder et al., 2021).

vii **Scout Neglect bias:** Important findings in preliminary or scout images may be overlooked because it is not expected to show meaningful pathology (Yoon et al., 2024).

2.2 Bias in Medical Imaging

RRG systems depend heavily on large datasets of medical images paired with corresponding radiology reports. Models inherit biases in medical imaging datasets, primarily demographic imbalances (e.g., racial or ethnic under-representation), reducing accuracy for minority groups (Koçak et al., 2025; Ricci Lara et al., 2022). We describe these biases as follows:

i **Sampling bias**: Sampling bias occurs when the dataset does not adequately represent the diversity of the patient population or disease spectrum. This phenomenon was explicitly noted in large vision–language models for chest X-rays: (Yang et al., 2025) found state-of-the-art models tended to underdiagnose pathologies in marginalized subgroups (e.g., black patients and especially black females) compared to radiologists. In that case, the AI model trained on this data may perform well on similar populations but poorly on underrepresented groups (Busby et al., 2018). For example, disparities have been observed in AI for detecting diabetic retinopathy (73% accuracy on light-skinned vs 60.5% on dark-skinned subjects) and for chest X-ray interpretation (higher rates of false negatives in underserved populations) (Ricci Lara et al., 2022).

ii **Annotation bias**: Annotation bias arises from variability and errors in how radiologists label images or write reports. Radiologists' subjective interpretations, experience levels, and cognitive biases can introduce systematic errors during annotation. For instance, radiologists may focus more on malignant lesions while under-annotating benign findings. This leads to skewed training data that causes AI models to detect specific pathologies over others preferentially (Catala et al., 2021; Yi et al., 2025a; Banerjee et al., 2023; Multusch et al., 2025).

iii **Propagation bias**: Propagation bias arises from data quality issues such as missing metadata, inconsistent report formats, transcription errors, and variations in imaging protocols across institutions. These issues introduce inconsistencies propagating through the AI training pipeline and may be amplified in the final model (Tripathi et al., 2023; Catala et al., 2021). For example, variability in imaging equipment or protocols across hospitals can result in institutional bias.

iv **Acquisition bias**: Images collected from different hospitals or devices may have systematic differences (e.g., scanner type, imaging protocol, resolution), resulting in acquisition bias (or domain shift). A model trained on high-quality research images might fail on routine clinical scans. For example, a model trained on 3T MRI scans may not generalize to the lower-resolution 1.5T scans commonly used in practice (Castro et al., 2020; Banerjee et al., 2023). Differences in scanner hardware or imaging parameters are well-known to cause distribution shifts in medical imaging. Addressing such bias is a significant challenge: performance can drop when applied to new clinical settings unless the model is explicitly trained to be invariant to acquisition factors.

2.3 Bias due to Large Language Models

Despite their strong report-generation capabilities, LLMs introduce several sources of bias in RRG, as discussed below.

i **Hallucinations**: LLMs sometimes generate plausible-sounding but factually incorrect statements called hallucinations (Huang et al., 2025; Tonmoy et al., 2024; Bang et al., 2023; Guerreiro et al., 2023). LLMs are known to hallucinate in RRG (Das et al., 2025; Nakaura et al., 2024b; Rahsepar et al., 2023). In the context of RRG, hallucination means describing a condition not present in the image, i.e., may generate fndings with content that cannot be directly linked to the input information, a serious danger in clinical use. In a recent review of LLMs for radiology reporting, all evaluated models (e.g., GPT-3.5, GPT-4 (Achiam et al., 2023), and fine-tuned variants) were found to hallucinate, with GPT-4 producing notably more false findings than a specialized vision-language model (Artsi et al., 2025; Tanno et al., 2025).

ii **Dataset imbalances and spurious correlations**: LLMs may learn spurious correlations between clinical findings and patient demographics or co-occurring diseases that do not reflect true causal relationships (Tanno et al., 2025). For example, (Voinea et al., 2024) found that fine-tuned Llama 3 on chest X-rays and noted that the model's conclusions lacked clinical judgment and exhibited biases due to dataset limitations.

iii **Sociodemographic bias**: LLMs also inherit biases from their training data (Yu et al., 2023). Pretrained on broad text corpora, they encode societal stereotypes and historical inequities (Shejole and Bhattacharyya, 2025; Shimabucoro et al., 2024; Nadeem et al., 2021; Nangia et al., 2020). In medical settings, these biases can manifest in subtle but harmful ways (Adiba et al., 2025; Omar et al., 2025). For instance, (Omiye et al., 2023) finds that GPT-4 could recapitulate debunked race-based medical misconceptions when answering questions. Similarly, (Yang et al., 2024) showed that GPT-3.5 and GPT-4 project higher costs and more extended hospitalizations for White populations and hold optimistic outcome views in harsh scenarios, reflecting real-world disparities. Although not specific to RRG, these studies show that LLM outputs can vary harmfully by patient demographics.

3 Causal Inference Perspective on RRG

Causal inference provides tools to analyze these biases by explicitly modeling cause–and–effect relationships. The necessary background on causal inference is detailed in Appendix J.

3.1 Causal Modeling of RRG

The generation of radiology reports can be viewed through causal modeling by constructing structural causal models (SCMs) that represent the relationships between imaging features, clinical variables, and textual descriptions. Many existing models inadvertently learn spurious correlations, such as associating standard anatomical features or frequent disease co-occurrences with specific report phrases, rather than the true causal factors. The research carried out by (Song et al., 2023; Jantscher et al., 2025; Vigneshwaran et al., 2024) explicitly model disease co-occurrences as confounders. They have observed that certain diseases often co-occur in the biased training data. Without accounting for this, an RRG model may learn spurious associations, always mentioning disease B whenever disease A is present, even if B is absent in the image. Bayesian networks and SCMs have actively used to discover causal associations between imaging findings and diseases from large-scale radiology report corpora (Ma et al., 2023; Pyrros et al., 2007; Do et al., 2017), achieving high precision in identifying true causal pairs. Moreover, causal modeling helps address dataset shift and annotation bias, such as when prior imaging or clinical history influences report content, by representing these factors as causal nodes and adjusting for their effects. Both approaches in-tandem improves the generalizability and clinical validity of generated reports.

RRG can be framed as a causal process involving multiple variables such as the medical image X, clinical context C, radiologist's interpretation I, and the final report text R. This process can be modeled using an SCM defined as a tuple $\mathcal{M} = \langle U, V, F \rangle$, where U is a set of exogenous variables (unobserved noise), $V = \{X, C, I, R\}$ is a set of endogenous variables, and $F = \{f_X, f_C, f_I, f_R\}$ denotes the set of causal mechanisms, where $X = f_X(U_X)$, $C = f_C(U_C)$, $I = f_I(X, C, U_I)$, and $R = f_R(I, U_R)$, where f_I captures how the image features and the clinical context influence the radiologist's interpretation, while f_R models how the interpretation generates the textual report. Biases emerge when confounders Z (e.g., patient demographics or annotation policies) influence both X and R, creating spurious associations and can be written as $X = f_X(U_X, Z), R = f_R(I, Z, U_R)$ To assess the causal effect of X on R, the do-operator is used, defining the interventional distribution as $P(R \mid \text{do}(X = x)) = \sum_z P(R \mid X = x, Z = z)P(Z = z)$.

Unknown confounders can jointly influence visual features (e.g., lung texture) and linguistic cues (report words). Conceptually, this mimics a front-door causal adjustment, where latent *mediators* are added to break the direct spurious path (Chen et al., 2023). Training with these modules encourages the model to rely on causal visual information. Constructing SCMs for RRG means including nodes for image features, diseases, patient attributes, and report tokens. By tracing the causal graph, one can identify sources of bias. The causal perspective forces transparency about assumptions: for instance, if it is assumed that demographic variables do not affect the diagnostic finding except via

disease prevalence, any direct edges from race/age to the report (bypassing disease) would indicate unfair bias (Castro et al., 2020).

3.2 Counterfactual Reasoning and Augmentation

Counterfactual reasoning asks how a model's output would change under a hypothetical intervention (Ji et al., 2023; Balashankar et al., 2021). Counterfactual methods have been used to generate augmented datasets and improve model robustness against bias (Song et al., 2023; Pitis et al., 2022; Uwaeze et al., 2025). In RRG, this approach facilitates the generation of counterfactual images by selectively altering critical regions (e.g., lung or heart patches) to simulate alternative diagnoses, thereby exposing and mitigating spurious correlations learned by models (Song et al., 2023). Formally, given an observed instance with variables $X = x$, $R = r$, the counterfactual outcome $R_{X=x'}$ represents the report that would have been generated had the image been x' instead of x. Using the SCM, the counterfactual is computed as $R_{X=x'}(u) = f_R\big(f_I(x', C(u), U_I(u)), U_R(u)\big)$ where u denotes a realization of all exogenous variables, counterfactual augmentation techniques generate synthetic samples by altering critical regions in X (e.g., lung patches) to x', producing new pairs (x', r') that help models learn invariant causal features. Frameworks like Counterfactual Feature Exchange (CoFE) (Li et al., 2024a) synthesize new image-report pairs by swapping lesion patches between positive and negative samples, effectively creating counterfactual images that differ only in the presence or absence of specific pathologies. This augmentation helps models learn to focus on causal features related to disease presence rather than confounding anatomical context. Similarly, counterfactual report reconstruction (Magic Cube) (Song et al., 2023) techniques generate alternative report narratives that exclude certain confounding disease mentions, breaking spurious co-occurrence patterns that commonly bias models.

Contrastive learning approaches further leverage counterfactual examples to teach models to distinguish between factual and counterfactual image representations, enhancing their generalization ability beyond training biases (Roschewitz et al., 2025; Li et al., 2024b; Aloui et al., 2023; Zhang et al., 2020; Shvetsov et al., 2024). These counterfactual strategies address critical challenges such as the "independence of diseases" problem, where models mistakenly infer causal relationships between unrelated conditions due to their frequent co-occurrence in the data. By explicitly training models on counterfactual variations, researchers can significantly reduce such biases and improve the clinical reliability of generated reports.

In addition to these image-based augmentations, text-level counterfactuals could be used. For example, one might imagine generating patient records with swapped demographic attributes to measure fairness. While not yet common in RRG, such methods have been used in fairness research (e.g., counterfactual examples where gender or race is changed (Howard et al., 2024; Sahoo et al., 2024; Nadeem et al., 2021; Nangia et al., 2020; Kusner et al., 2017; Mehta et al., 2026)) and could be applied to modify image–report pairs. Overall, counterfactual reasoning concretely implemented as data synthesis or perturbation is a key causal tool for RRG bias mitigation.

3.3 Causal Debiasing in LLMs

Causal ideas have been applied to mitigate language-model biases in NLG tasks (Sun et al., 2024; Wu et al., 2024; Zhou et al., 2023). Although most studies focus on social or linguistic biases, similar principles apply to medical text generation. Causal debiasing methods for LLMs aim to regulate these biases by incorporating causal reasoning into the generation process. Causal prompting (Zhang et al., 2025) is one such front-door adjustment approach. It involves altering the input prompt to steer the LLM away from undesirable biases. It treats the chain-of-thought generated by the LLM as a mediator variable, and then computes the causal effect of the prompt on the answer by marginalizing over the chain-of-thought. Causal debiasing techniques integrate causal interventions such as back-door and front-door adjustments to remove confounding effects by redesigning prompts (without changing model parameters) to simulate intervening on hidden reasoning. See Appendix J.4 for details on front- and back-door adjustments.

Another approach is causality-guided selection and filtering. Li et al. (2025a) outlines a framework called Prompting Fairness, which first identifies how social information (e.g., terms related to race or gender) flows through an LLM's reasoning graph, and then applies selection mechanisms in the prompt to block or weaken biased paths. Beyond prompting, one can use adversarial debiasing in representation learning for instance, training

Table 1: Summary of Bias in RRG with Mitigation Ability of Causal Inference Techniques.

Bias Category	Bias	Mitigable via Causal Inference
Linguistic	Alliterative	No
Heuristic	Framing	Yes
	Availability	
	Anchoring	No
Heuristic (Domain-specific)	Scout Neglect	Partially
Cognitive	Hindsight	Yes
	Blind Spot	No
Data-related	Sampling Bias	Yes
	Annotation Bias	Partially
	Acquisition Bias	Yes
	Data Imbalance	Yes
Systemic	Propagation Bias	Partially
Model-related	LLM Hallucination	Partially
Societal	Sociodemographic Bias	Partially

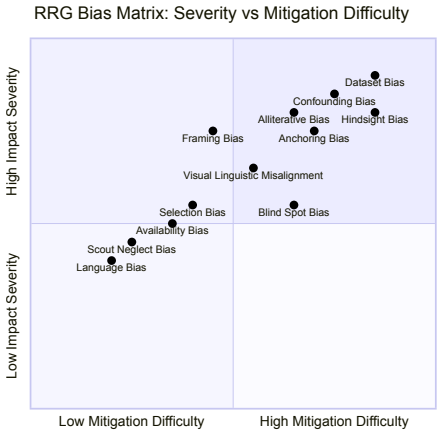

Figure 2: RRG Bias Matrix describing Severity vs Mitigation difficulty

the LLM (or its adapter) with an adversary trying to predict patient attributes from its hidden states, thereby encouraging invariant representations.

In practice, learnable prompts (Moore et al., 2024; Lester et al., 2021; Li and Liang, 2021; Elazar et al., 2021) encoding factual and counterfactual information act as interventions that guide LLMs to generate semantically coherent and factually accurate reports. The training objective incorporates these causal adjustments (e.g., by minimizing a loss function weighted to penalize predictions correlated with confounders).

Table 1 summarizes biases in RRG and the extent to which causal inference mitigates them (see Appendix D for details). Figure 2 maps bias types along severity and mitigation difficulty, revealing stratification by clinical risk and tractability (Appendix E). Appendix F outlines the causal inference pipeline in RRG, showing how factors affect image acquisition, representation learning, report generation, and evaluation, leading to distinct biases. A structured causal view of bias interrelationships is provided in Appendix G.

4 Categorization of Mitigation Strategies

We broadly categorize various interventions against bias in RRG by their application stage as data-level, model-level, and evaluation-level intervention. Each category contains specific techniques, often inspired by causal thinking.

4.1 Data-level intervention

Data-level intervention focuses on improving the quality and representativeness of training data to reduce confounding and selection biases inherent in radiology datasets. It invoves adjusting the dataset to equalize representation. For example, if certain patient subgroups or rare pathologies are underrepresented, one can oversample those cases or undersample overrepresented ones (Koçak et al., 2025). Causal sampling can also create a balanced distribution conditional on key factors, mimicking an intervention that breaks confounding. New data can be generated that explicitly decorrelates confounders, which can be done by synthesizing images with or without a particular finding (e.g., using generative models or patch-mixing) to break spurious co-occurrences (Song et al., 2023; Li et al., 2024a). Domain knowledge can augment datasets with rare or critical cases; for example, adding diverse COVID-19 or tuberculosis chest X-rays can reduce bias toward common Western diseases (Wynants et al., 2020; Zech et al., 2018; Oakden-Rayner et al., 2020). Data can be re-annotated to reduce label bias. For instance, correcting systematic errors in report labels or using multiple annotators for ambiguous cases can help produce a fairer ground truth (Santomartino et al., 2024).

4.2 Model-level intervention

Model-level interventions incorporate causal principles directly into the architecture and training objectives of RRG systems to disentangle true disease signals from confounders. For instance, the Visual-Linguistic Causal Intervention (VLCI) framework (Chen et al., 2023) employs visual and linguistic deconfounding modules that implicitly mitigate cross-modal confounders by applying causal front-door interventions. This approach helps the model focus on medically relevant features rather than superficial correlations such as high-frequency con-

text words or salient but irrelevant visual patterns. When using large LLM decoders, one can fine-tune or prompt them in a debiasing way. For instance, the causal prompting methods (Li et al., 2024a, 2025a) are model-level since they alter input-output processing. Such causal debiasing ensures that language models generate reports grounded in actual pathology rather than dataset artifacts, improving diagnostic accuracy and trustworthiness.

4.3 Evaluation-level intervention

Evaluation-level interventions do not change the model but change how we measure or enforce fairness during testing and deployment. By understanding the causal pathways that generate observed data, evaluation frameworks can better assess whether models capture true disease mechanisms and avoid misleading correlations (Baradwaj et al., 2024). Counterfactual simulations (e.g., changing patient age) can reveal biases if model outputs shift without explanation. Post-processing and interpretability techniques further identify residual biases and help clinicians interpret decisions causally (Kibria et al., 2025; Jiao et al., 2025; Zamir et al., 2025), ensuring generated reports are clinically valid. Causal inference tools should be applied at evaluation, for example by estimating the average causal effect of protected attributes on model outputs via counterfactual sampling to quantify unintended influence. Ultimately, radiologist review remains essential, as targeted audits focused on known biases, such as sampling underrepresented groups or checking for common hallucinations, can reveal failures missed by automated metrics. Over time, audit feedback can guide retraining or redesign, helping address dataset shift, selection bias, and spurious correlations that limit RRG models.

5 Analyzing Datasets and Metrics

5.1 RRG Datasets

Table 2 summarizes widely used public medical imaging benchmarks that exhibit age and geographic biases due to non-random data collection, limited geographic coverage, and incomplete demographic annotation. Most datasets originate from a few tertiary-care institutions in high-income countries, inducing strong selection effects that skew age and disease severity. Age distributions often misalign with real-world prevalence, confounding demographics with disease labels, while race and ethnicity metadata are frequently missing or inconsistently reported. From a causal perspective, demographic and geographic biases arise from three mechanisms: confounding (demographics affect both disease prevalence and image appearance), mediation (demographics influence healthcare processes that alter images), and selection bias (dataset inclusion depends on institutional or clinical pathways). Ignoring these mechanisms undermines model interpretability and clinical reliability. Broader demographic representation (e.g., age and geographic diversity) should be incorporated into datasets to enhance their global applicability.

Table 2: Summary of age and geographic biases in commonly used medical imaging datasets.

Organ	Datasets	Geographic Bias	Age Bias
Chest	MIMIC-CXR (Johnson et al., 2019)	United States (Boston)	Predominantly elderly ICU patients
	IU-Xray (Demner-Fushman et al., 2015)	United States (Indiana University)	Middle-aged (mean age $\approx$ 49 years)
	CheXpert (Irvin et al., 2019)	United States (Stanford University)	Middle-aged adults
	PadChest (Bustos et al., 2020)	Spain	Elderly patients
Spine	OAI (Peterfy et al., 2008), RSNA Spine (Flanders et al., 2022)	Germany, United States	Degenerative cases overrepresented
Abdomen	LiTS (Bilic et al., 2023)	United States	Middle-aged adults
	KiTS (Heller et al., 2019)	United States (University of Minnesota Medical Center)	
	CHAOS (Kavur et al., 2021)	China	
Brain	BraTS (Menze et al., 2015), CQ500 (Chilamkurthy et al., 2018)	India, United States, China	Adult glioma cases (age $\geq$ 35 years)

5.2 Evaluation Metrics

Evaluation metrics in RRG such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015) that measure textual overlap but are agnostic to causal structure. As a result, these metrics cannot detect or correct biases arising from confounding, mediation, or selection bias in the data. In contrast, causal-aware evaluation metrics (Table 3) align model assessment with clinically meaningful variables by corresponding to interventional or counterfactual queries, making bias mitigation identifiable and actionable. Concept-level evaluation metrics such as CheXbert F1 (Irvin et al., 2019) and RadGraph F1 (Jain et al., 2021) are highly sensitive to spurious correlations in radiology data, for example when patient in Intensive care unit (ICU) is most likely to be associated with medical devices

Table 3: Causal-aware metrics applicable in RRG. F is findings, X is input, E, R are entity relationships, Y is outcome, A is variable name, and $\mathbb{E}(\cdot)$ is expectation.

Metrics	Causal Variable Evaluated	Causal Estimand	Bias Type Detected
CheXpertF1 (Irvin et al., 2019)	Clinical findings (disease presence)	$P(F\|\text{do}(X))$	Confounding, annotation bias
RadGraph F1 (Jain et al., 2021)	Entities & relations	$P(E,R\|\text{do}(X))$	Mediation
Demographic Performance Gap (Obermeyer et al., 2019)	Group-conditioned findings	$\mathbb{E}[Y\,\|\,\text{do}(D{=}g)]$	Demographic confounding
Calibration Error (ECE/Brier) (Guo et al., 2017; Subbaswamy and Saria, 2019)	Model confidence vs correctness	$P(Y\|\text{do}(X))$	Selection bias, dataset shift
Counterfactual Consistency (Pearl et al., 2016)	Stability under irrelevant changes	$Y_{A\leftarrow a} = Y_{A\leftarrow a'}$	Shortcut learning
Invariant Risk / Stability Score (Arjovsky et al., 2019; Gulrajani and Lopez-Paz, 2021; Magliacane et al., 2018)	Performance across environments	$Y \perp E\|\text{do}(X)$	Dataset shift
Transportability Error (Castro et al., 2020; Bareinboim and Pearl, 2013)	Cross-dataset correctness	$P(Y\|\text{do}(X))$	Institutional bias

(e.g., tubes) that are themselves correlated with disease labels. As a result, high metric scores may reflect shortcuts rather than true clinical reasoning. To address this, these metrics can be evaluated conditional on key confounders such as age, sex, and scanner type, and compared under causal interventions (e.g., do(ICU = outpatient)). In addition, the Demographic Performance Gap (DAP) measures accuracy differences across demographic groups, helping to identify demographic confounding, fairness violations, and counterfactual unfairness in report generation systems.

Calibration error (e.g., ECE, Brier score) (Guo et al., 2017; Subbaswamy and Saria, 2019) assesses whether predicted confidence matches true correctness; under dataset shift, miscalibration indicates selection bias arising from changes in the data-generating process. Counterfactual consistency (Pearl et al., 2016) tests invariance of generated reports to non-causal attributes (e.g., sex, care setting) given fixed pathology, directly reflecting counterfactual queries. Invariant risk minimization (IRM) (Arjovsky et al., 2019) aims to learn image–text generation models whose performance remain stable across different radiological imaging environments. Using IRM (Gulrajani and Lopez-Paz, 2021) and causally grounded representation-learning methods (Magliacane et al., 2018) can reduce shortcut-driven hallucinations. Transportability error (Bareinboim and Pearl, 2013) quantifies performance drop across datasets (e.g., MIMIC-CXR to IU-Xray), capturing institutional and acquisition biases that limit external validity.

6 Conclusion and Future Directions

Fairness in RRG is an important concern as medical AI tools move toward clinical use. In this survey, we reviewed the origins of bias in the RRG pipeline. Since these biases mainly arise from spurious correlations, examining cause-effect relationships can help mitigate them. Accordingly, we modeled RRG as a causal model and explored causal inference techniques such as counterfactuals and causal debiasing. To assist researchers in selecting and applying appropriate bias mitigation methods, we categorized mitigation approaches into a three-tier taxonomy based on the level of intervention. We also analyzed which biases are mitigable and non-mitigable using causal inference. Further, we examined popular medical imaging datasets and identified various biases within them. We analyzed current evaluation metrics through a causal lens and found that they fail to capture causal alignment and clinical fidelity. We argued that broader demographic coverage in datasets (e.g., age and geographic diversity) and the development of causal metrics can improve fairness and reliability in RRG. Continued research integrating causal reasoning, domain expertise, and advanced machine learning is essential for advancing trustworthy medical imaging AI, making this survey a practical guide for researchers and clinicians developing radiology AI systems.

Future work should prioritize causal fairness criteria, such as evaluating whether report content remains invariant under counterfactual changes to sensitive attributes while pathology is held fixed. High-priority directions include developing counterfactual and group-conditional evaluation metrics aligned with causal estimands, as well as causal auditing and intervention strategies for foundation models, including targeted counterfactual testing and controlled fine- tuning, to ensure that performance gains do not compromise fairness or clinical validity. Additional details and guidelines for clinicians and researchers are provided in Appendices H and I, respectively.

Limitations

This survey focused on bias obtained from the published research at the intersection of RRG, fairness, and causal inference; many practical systems or proprietary models are not publicly documented. Thus, real-world deployment issues (regulatory constraints, liability, workflow integration) are only briefly mentioned. The emphasis was more on technical methods and did not comprehensively cover sociotechnical factors. Topics like patient privacy laws, the cost of collecting balanced datasets, and the ethics of AI in radiology were largely beyond our scope. RRG is closely related to other medical language tasks (like report summarization or question answering) where causal bias methods may also apply. Still, these adjacent areas are not explored in detail as our focus was specifically on RRG.

Ethical Considerations

RRG systems operate in a high-stakes clinical domain, where errors or biases in generated reports can influence diagnostic reasoning, decision-making, and patient diagnosis. Ethical considerations in surveying causal methods, bias mitigation strategies, and evaluation protocols for RRG extend beyond general concerns in natural language generation (NLG) and must be grounded in clinical safety, fairness, and accountability. A primary ethical concern is the risk of reinforcing existing health disparities; for instance, models trained on imbalanced data may underdiagnose pathologies in marginalized subgroups, such as Black female patients, or recapitulate debunked race-based medical misconceptions. To mitigate these risks, we advocate for a causal inference perspective that prioritizes the identification of true underlying disease mechanisms over spurious correlations. We emphasize that RRG systems should not be deployed as autonomous diagnostic tools but as supportive aids within a human-in-the-loop framework, requiring rigorous validation via causal-aware metrics and prospective clinical trials to ensure that performance gains do not come at the cost of fairness or clinical validity.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farzana Islam Adiba, Yifan Zhang, and Rahmatollah Beheshti. 2025. Bias and fairness in medical llms: An extensive scoping review. *OSF*.

Sohrab Afshari Mirak, Sree Harsha Tirumani, Nikhil Ramaiya, and Inas Mohamed. 2025. The growing nationwide radiologist shortage: current opportunities and ongoing challenges for international medical graduate radiologists. *Radiology*, 314(3):e232625.

Ahmed Aloui, Juncheng Dong, Cat P Le, and Vahid Tarokh. 2023. Counterfactual data augmentation with contrastive learning. *arXiv preprint arXiv:2311.03630*.

Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Richa Arora. 2014. The training and practice of radiology in india: current trends. *Quantitative imaging in medicine and surgery*, 4(6):449.

Yaara Artsi, Eyal Klang, Jeremy D. Collins, Benjamin S. Glicksberg, Panagiotis Korfiatis, Girish N Nadkarni, and Vera Sorin. 2025. Large language models in radiology reporting—a systematic review of performance, limitations, and clinical implications. *medRxiv*.

Ananth Balashankar, Xuezhi Wang, Ben Packer, Nithum Thain, Ed Chi, and Alex Beutel. 2021. Can we improve model robustness through secondary attribute counterfactuals? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4701–4712.

Imon Banerjee, Kamanasish Bhattacharjee, John L. Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N. Patel, Rakesh Shiradkar, and Judy Gichoya. 2023. "shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9):842–851.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Simha Sankar Baradwaj, Destiny Gilliland, Jack Rincon, Henning Hermjakob, Yu Yan, Irsyad Adam, Gwyneth Lemaster, Dean Wang, Karol Watson, Alex Bui, et al. 2024. Building an ethical and trustworthy biomedical ai ecosystem for the translational and clinical integration of foundational models. *arXiv preprint arXiv:2408.01431*.

Elias Bareinboim and Judea Pearl. 2013. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Rajesh Bhayana. 2024. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*, 310(1):e232756.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. 2023. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680.

Lindsay P Busby, Jesse L Courtier, and Christine M Glastonbury. 2018. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, 38(1):236–247.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, et al. 2025. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, 35(5):2589–2602.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Chelsea Castillo, Tom Steffens, Lawrence Sim, and Liam Caffery. 2021. The effect of clinical information on radiology reporting: a systematic review. *Journal of medical radiation sciences*, 68(1):60–74.

Daniel C. Castro, Ian Walker, and Ben Glocker. 2020. Causality matters in medical imaging. *Nature Communications*, 11(1):3673.

Omar Del Tejo Catala, Ismael Salvador Igual, Francisco Javier Perez-Benito, David Millan Escriva, Vicent Ortiz Castello, Rafael Llobet, and Juan-Carlos Perez-Cortes. 2021. Bias analysis on public x-ray image datasets of pneumonia and COVID-19 patients. *IEEE Access*, 9:42370–42383.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Shen Zhao, Guanbin Li, Cheng-Lin Liu, and Liang Lin. 2023. Cross-modal causal intervention for medical report generation. *arXiv preprint arXiv:2303.09117*.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. Development and validation of deep learning algorithms for detection of critical findings in head ct scans. *arXiv preprint arXiv:1803.05854*.

Anindya Bijoy Das, Shahnewaz Karim Sakib, and Shibbir Ahmed. 2025. Trustworthy medical imaging with large language models: A study of hallucinations across modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1265–1272.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Bao H. Do, Curtis Langlotz, and Christopher F. Beaulieu. 2017. Bone tumor diagnosis using a naïve bayesian model of demographic and radiographic features. *Journal of Digital Imaging*, 30(5):640–647.

Kyung-Hyun Do, Kyongmin Sarah Beck, and Jeong Min Lee. 2023. The growing problem of radiologist shortages: Korean perspective. *Korean Journal of Radiology*, 24(12):1173.

Yanai Elazar et al. 2021. Measuring and improving model-moderated counterfactual reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 997–1011.

Adam Flanders, Chris Carr, Errol Colak, PhD FelipeKitamura, MD, Hui Ming Lin, JeffRudie, John Mongan, Katherine Andriole, Luciano Prevedello, Michelle Riopel, Robyn Ball, and Sohier Dane. 2022. Rsna 2022 cervical spine fracture detection. https://kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection. Kaggle.

FM Grieve, AA Plumb, and SH Khan. 2010. Radiology reporting: a general practitioner's perspective. *The British journal of radiology*, 83(985):17–22.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Linda L Humphrey, Benjamin KS Chan, and Harold C Sox. 2002. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of internal medicine*, 137(4):273–284.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Jason N. Itri and Sohil H. Patel. 2018. Heuristics and cognitive error in medical imaging. *American Journal of Roentgenology*, 210(5):1097–1105. PMID: 29528716.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Michael Jantscher, Felix Gunzer, Gernot Reishofer, and Roman Kern. 2025. Causal insights from clinical information in radiology: Enhancing future multimodal ai development. *Computer Methods and Programs in Biomedicine*, 268:108810.

Jianchao Ji, Zelong Li, Shuyuan Xu, Max Xiong, Juntao Tan, Yingqiang Ge, Hao Wang, and Yongfeng Zhang. 2023. Counterfactual collaborative reasoning. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 249–257.

Wenpei Jiao, Kun Shang, Hui Li, Ke Yan, Jiajin Zhang, Guangjie Yang, Lijuan Guo, Yan Wan, Xing Yang, Dakai Jin, et al. 2025. Vision-language models for automated 3d pet/ct report generation. *arXiv preprint arXiv:2511.20145*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty, and Rustom Lawyer. 2023a. Kgvlbart: knowledge graph augmented visual language bart for radiology report generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3401–3411.

Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Jadhav. 2023b. Replace and report: Nlp assisted radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10731–10742.

Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Milind Gune, Kush Shrivastava, Rustom Lawyer, and Spriha Biswas. 2022. Knowledge enhanced deep learning model for radiology text generation. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 32–42.

Navdeep Kaur, Ajay Mittal, and Gurprem Singh. 2022. Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey. *Multimedia Tools and Applications*, 81(10):13409–13439.

A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. 2021. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950.

Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. The pursuit of fairness in artificial intelligence models: A survey. *arXiv preprint arXiv:2403.17333*.

Md Raisul Kibria, Sébastien Lafond, and Janan Arslan. 2025. Decoding the multimodal maze: A systematic review on the adoption of explainability in multimodal attention-based models. *arXiv preprint arXiv:2508.04427*.

Sunkyu Kim, Choong-kun Lee, and Seung-seob Kim. 2024. Large language models: a guide for radiologists. *Korean Journal of Radiology*, 25(2):126.

Burak Koçak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E Klontzas, Roberto Cannella, and Renato Cuocolo. 2025. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagn. Interv. Radiol.*, 31(2):75–88.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Cindy S. Lee, Paul G. Nagy, Sallie J. Weaver, and David E. Newman-Toker. 2013. Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology*, 201(3):611–617. PMID: 23971454.

Ryan C Lee, Roham Hadidchi, Michael C Coard, Yossef Rubinov, Tharun Alamuri, Aliena Liaw, Rahul Chandrupatla, and Tim Q Duong. 2025. Use of large language models on radiology reports: A scoping review. *Journal of the American College of Radiology*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2025a. Prompting fairness: Integrating causality to debias large language models. In *The Thirteenth International Conference on Learning Representations*.

Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. 2024a. Contrastive learning with counterfactual explanations for radiology report generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLIII*, page 162–180, Berlin, Heidelberg. Springer-Verlag.

Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. 2024b. Contrastive learning with counterfactual explanations for radiology report generation. In *European Conference on Computer Vision*, pages 162–180. Springer.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4582–4597.

Yilin Li, Chao Kong, Guosheng Zhao, and Zijian Zhao. 2025b. Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. *Artificial Intelligence Review*, 58(11):1–42.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chang Liu, Yuanhe Tian, and Yan Song. 2023. A systematic review of deep learning-based research on radiology report generation. *arXiv preprint arXiv:2311.14199*.

R. Liu, C. Shi, Regan Song, M. Niethammer, T. Li, and H. Zhu. 2025. Hcdpd: A heterogeneous causal framework for disease pattern detection in medical imaging. *medRxiv : the preprint server for health sciences*.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.

Peiqing Lv, Yaonan Wang, Min Liu, Zhe Zhang, Yunfeng Ma, Licheng Liu, and Erik Meijering. 2025. CiSeg: Unsupervised cross-modality adaptation for 3D medical image segmentation via causal intervention. *IEEE Trans. Med. Imaging*, PP:1–1.

Shawn X. Ma, Ali H. Dhanaliwala, Jeffrey D. Rudie, Andreas M. Rauschecker, Douglas Roberts-Wolfe, Peter Haddawy, and Charles E. Kahn. 2023. Bayesian networks in radiology. *Radiology: Artificial Intelligence*, 5(6):e210187.

Salvatore Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Joris M Mooij, Bernhard Schölkopf, et al. 2018. Domain generalization via invariant feature representation. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Dima Mamdouh, Mariam Attia, Mohamed Osama, Nesma Mohamed, Abdelrahman Lotfy, Tamer Arafa, Essam A Rashed, and Ghada Khoriba. 2025. Advancements in radiology report generation: A comprehensive analysis. *Bioengineering*, 12(7):693.

Luis Martí-Bonmatí. 2021. Estimates of causality with medical image in oncology. *ANALES RANM*, 138:16–23.

Raghav Mehta, Fabio De Sousa Ribeiro, Tian Xia, Mélanie Roschewitz, Ainkaran Santhirasekaram, Dominic C. Marshall, and Ben Glocker. 2026. Cf-seg: Counterfactuals meet segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 117–127, Cham. Springer Nature Switzerland.

Patricio Melendez-Rojas, Jaime Jamett-Rojas, María Fernanda Villalobos-Dellafiori, Pablo R Moya, and Alejandro Veloz-Baeza. 2025. Current landscape of automatic radiology report generation with deep learning: An exploratory systematic review.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia,

Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. 2015. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878.

Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. 2024. Reasoning beyond bias: A study on counterfactual prompting and chain of thought reasoning. *arXiv preprint arXiv:2408.08651*.

Malte Michel Multusch, Lasse Hansen, Mattias Paul Heinrich, Lennart Berkel, Axel Saalbach, Heinrich Schulz, Franz Wegner, Joerg Barkhausen, and Malte Maria Sieren. 2025. Impact of radiologist experience on AI annotation quality in chest radiographs: A comparative analysis. *Diagnostics (Basel)*, 15(6):777.

A. Murphy, F. Dixon, and F. Deng. 2024. Cognitive bias in diagnostic radiology. https://radiopaedia.org/articles/cognitive-bias-in-diagnostic-radiology?lang=us. Accessed: 2025-04-12.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371.

Takeshi Nakaura, Rintaro Ito, Daiju Ueda, Taiki Nozaki, Yasutaka Fushimi, Yusuke Matsui, Masahiro Yanagawa, Akira Yamada, Takahiro Tsuboyama, Noriyuki Fujima, et al. 2024a. The impact of large language models on radiology: a guide for radiologists on the latest innovations in ai. *Japanese journal of radiology*, 42(7):685–696.

Takeshi Nakaura, Naofumi Yoshida, Naoki Kobayashi, Kaori Shiraishi, Yasunori Nagayama, Hiroyuki Uetani, Masafumi Kidoh, Masamichi Hokamura, Yoshinori Funama, and Toshinori Hirai. 2024b. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, 42(2):190–200.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1953–1967.

Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. 2023. Pragmatic radiology report generation. In *Machine Learning for Health (ML4H)*, pages 385–402. PMLR.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. 2025. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, pages 1–9.

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1):195.

Omer Onder, Yasin Yarasir, Aynur Azizova, Gamze Durhan, Mehmet Ruhi Onur, and Orhan Macit Ariyurek. 2021. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights into Imaging*, 12(1):51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causaf Inference in Statistics: A Primer*. Wiley.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Derek C. Penn and Daniel J. Povinelli. 2007. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology*, 58(Volume 58, 2007):97–118.

C G Peterfy, E Schneider, and M Nevitt. 2008. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis Cartilage*, 16(12):1433–1441.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. 2022. Mocoda: Model-based counterfactual data augmentation. *Advances in Neural Information Processing Systems*, 35:18143–18156.

Ayis Pyrros, Paul Nikolaidis, Vahid Yaghmai, Steve Zivin, Joseph I. Tracy, and Adam Flanders. 2007. A bayesian approach for the categorization of radiology reports. *Academic Radiology*, 14(4):426–430.

Amir Ali Rahsepar, Neda Tavakoli, Grace Hyun J Kim, Cameron Hassani, Fereidoun Abtin, and Arash Bedayat. 2023. How ai responds to common lung cancer questions: Chatgpt versus google bard. *Radiology*, 307(5):e230922.

C Rainey, A England, PC Murphy, AA Mohammad, YH Hadi, and M McEntee. 2026. Large language models (llms) in radiography research: A narrative review. *Radiography*, 32(1):103244.

Graciela Ramirez-Alonso, Olanda Prieto-Ordaz, Roberto López-Santillan, and Manuel Montes-Y-Gómez. 2022. Medical report generation through radiology images: an overview. *IEEE Latin America Transactions*, 20(6):986–999.

Bruce I Reiner, Nancy Knight, and Eliot L Siegel. 2007. Radiology reporting, past, present, and future: the radiologist's perspective. *Journal of the American College of Radiology*, 4(5):313–319.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. 2022. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1):4581.

Abi Rimmer. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359.

Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. 2025. Robust image representations with counterfactual contrastive learning. *Medical Image Analysis*, page 103668.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. Indibias: A benchmark dataset to measure social biases in language models for indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806.

Samantha M Santomartino, John R Zech, Kent Hall, Jean Jeudy, Vishwa Parekh, and Paul H Yi. 2024. Evaluating the performance and bias of natural language processing tools in labeling chest radiograph reports. *Radiology*, 313(1):e232746.

Jarrel CY Seah, Jennifer SN Tang, and Aengus Tran. 2025. Drafting the future: the dawn of ai report generation in radiology. *Radiology*, 316(1):e243378.

Kaustubh Shivshankar Shejole and Pushpak Bhattacharyya. 2025. Stereodetect: Detecting stereotypes and anti-stereotypes the correct way using social psychological underpinnings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4051–4082.

Jingpu Shi and Beau Norgeot. 2022. Learning causal effects from observational data in healthcare: A review and summary. *Frontiers in Medicine*, Volume 9 - 2022.

Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. Llm see, llm do: Leveraging active inheritance to target non-differentiable objectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9243–9267.

Dmytro Shvetsov, Joonas Ariva, Marharyta Domnich, Raul Vicente, and Dmytro Fishman. 2024. Coin: Counterfactual inpainting for weakly supervised semantic segmentation for medical images. In *Explainable Artificial Intelligence*, pages 39–59, Cham. Springer Nature Switzerland.

Xiao Song, Jiafan Liu, Yan Liu, Yun Li, Wenbin Lei, and Ruxin Wang. 2023. Rethinking radiology report generation via causal inspired counterfactual augmentation. In *ACM International Conference on Bioinformatics, Computational Biology and Biomedicine*.

Adarsh Subbaswamy and Suchi Saria. 2019. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352.

Zhouhao Sun, Li Du, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024. Causal-guided active learning for debiasing large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14455–14469.

Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, Karan Singhal, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Sara Mahdavi, Zahra Ahmed, Yossi Matias, Joelle Barral, S. M. Ali Eslami, Danielle Belgrave, Yun Liu, Sreenivasa Raju Kalidindi, Shravya Shetty, Vivek Natarajan, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. 2025. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608.

Ali S. Tejani, Yee Seng Ng, Yin Xi, and Jesse C. Rayan. 2024. Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics*, 44(5):e230067. PMID: 38635456.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Satvik Tripathi, Kyla Gabriel, Suhani Dheer, Aastha Parajuli, Alisha Isabelle Augustin, Ameena Elahi, Omar Awan, and Farouk Dako. 2023. Understanding biases and disparities in radiology ai datasets: A review. *Journal of the American College of Radiology*, 20(9):836–841.

Jason Uwaeze, Pranav Kulkarni, Vladimir Braverman, Michael A. Jacobs, and Vishwa S. Parekh. 2025. Generative counterfactual augmentation for bias mitigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1153–1160.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Vibujithan Vigneshwaran, Erik Ohara, Matthias Wilms, and Nils Forkert. 2024. Macaw: a causal generative model for medical imaging. *arXiv preprint arXiv:2412.02900*.

Ștefan-Vlad Voinea, Mădălin Mămuleanu, Rossy Vlăduț Teică, Lucian Mihai Florescu, Dan Selișteanu, and Ioana Andreea Gheonea. 2024. Gpt-driven radiology report generation with fine-tuned llama 3. *Bioengineering*, 11(10):1043.

Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. 2024. A survey of deep learning-based radiology report generation using multimodal data. *arXiv preprint arXiv:2405.12833*.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Laure Wynants, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, et al. 2020. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. *BMJ*, 369.

Zibo Xu, Qiang Li, Wei zhi Nie, Weijie Wang, and Anan Liu. 2025. Structure causal models and llms integration in medical visual question answering. *IEEE Transactions on Medical Imaging*, 44:3476–3489.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(1):176.

Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J Wang, Dushyant Sahani, and Shwetak Patel. 2025. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305.

Paul H. Yi, Preetham Bachina, Beepul Bharti, Sean P. Garin, Adway Kanhere, Pranav Kulkarni, David Li, Vishwa S. Parekh, Samantha M. Santomartino, Linda Moy, and Jeremias Sulam. 2025a. Pitfalls and best practices in evaluation of ai algorithmic biases in radiology. *Radiology*, 315(2):e241674.

Ziruo Yi, Ting Xiao, and Mark V Albert. 2025b. A survey on multimodal large language models in radiology for report generation and visual question answering. *Information*, 16(2):136.

Se-Young Yoon, Karen S. Lee, Abraham F. Bezuidenhout, and Jonathan B. Kruskal. 2024. Spectrum of cognitive biases in diagnostic radiology. *RadioGraphics*, 44(7):e230059.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36:55734–55784.

Muhammad Tayyab Zamir, Safir Ullah Khan, Alexander Gelbukh, Edgardo Manuel Felipe Riverón, and Irina Gelbukh. 2025. Explainable ai-driven analysis of radiology reports using text and image data: Experimental study. *JMIR Form Res*, 9:e77482.

John R. Zech, Marcus A. Badgeley, Mia Liu, Antonio B. Costa, Joseph J. Titano, and Eric K. Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLOS Medicine*, 15(11).

Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025. Causal prompting: Debiasing large language model prompting based on front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25842–25850.

Li Zhang, Xin Wen, Jian-Wei Li, Xu Jiang, Xian-Feng Yang, and Meng Li. 2023. Diagnostic error and bias in the department of radiology: a pictorial essay. *Insights into Imaging*, 14(1):163.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241.

A Taxonomy Chart

Unified taxonomy provides a conceptual framework (Figure 3) for understanding RRG through a causal lens, organizing how biases emerge across the RRG pipeline and how causal interventions can be systematically aligned with mitigation and evaluation.

B Prior Surveys in RRG

As discussed in the introduction (§ 1), a systematic analysis of prior RRG surveys reveals a critical research gap in the integration of causal inference. In this section, we provide more details about prior surveys in RRG and discuss how there is a critical need of integrating causal inference.

Early perspectives on radiology reporting, grounded in radiologists' and referring clinicians' needs, emphasised clarity, clinical relevance, and contextual reasoning in reports, long before the advent of deep learning models (Reiner et al., 2007; Grieve et al., 2010). These works highlighted that reporting is not merely a transcription of visual findings, but a reasoning process shaped by clinical context, prior knowledge, and diagnostic intent—an insight that remains highly relevant to modern AI-driven systems.

With the rise of deep learning, multiple surveys have systematically reviewed automatic RRG methods. Foundational surveys focus on convolutional and recurrent architectures, dataset characteristics, and evaluation metrics (Monshi et al., 2020; Kaur et al., 2022; Ramirez-Alonso et al., 2022). More recent reviews provide comprehensive taxonomies of multimodal pipelines, covering dataset availability and adoption, encoder–decoder designs, attention mechanisms, transformer-based models, and training strategies such as contrastive learning and reinforcement learning (Wang et al., 2024; Li et al., 2025b; Melendez-Rojas et al., 2025).

Several works have extended the scope of using deep learning architectures by examining clinical knowledge incorporation. (Wang et al., 2024; Yi et al., 2025b) have discussed the use of structured clinical data, multimodal fusion, and knowledge graphs to enhance report completeness and clinical fidelity, whereas other surveys (Nguyen et al., 2023; Seah et al., 2025) have emphasised pragmatic considerations, such as deployment constraints, robustness, and alignment with radiology workflows . Evaluation practices are also critically reviewed, highlighting the dominance of NLP similarity metrics (e.g., BLEU, ROUGE, CIDEr) alongside emerging qualitative clinical assessments, while noting persistent limitations in capturing true clinical correctness (Monshi et al., 2020; Li et al., 2025b).

With the rapid adoption of large language models (LLMs), their application in radiology includes spanning report drafting, structured reporting, question answering, and decision support has also seen an increasing growth in surveys carried out in this field of radiology (Bhayana, 2024; Nakaura et al., 2024a; Kim et al., 2024; Lee et al., 2025; Busch et al., 2025; Mamdouh et al., 2025; Rainey et al., 2026). These reviews highlight the transformative potential of LLMs, particularly when combined with vision models in multimodal large language model (MLLM) frameworks (Yi et al., 2025b). However, they primarily frame progress in terms of scale, fluency, and alignment with radiology-specific tasks, rather than underlying causal reasoning. Despite their breadth, existing surveys predominantly adopt a correlational perspective. Models are typically evaluated on their ability to reproduce report text patterns given image–report pairs, implicitly assuming that learning statistical associations is sufficient for reliable clinical reporting. Yet, systematic reviews on the effect of clinical information demonstrate that radiology reporting is causally influenced by prior history, indications, and contextual factors (Castillo et al., 2021). While current RRG surveys acknowledge multimodal inputs, they rarely examine whether models meaningfully reason about cause–effect relationships linking imaging findings, clinical context, and diagnostic conclusions.

As a result, a critical gap remains no survey to date systematically examines RRG through a causal lens. There remains a foundational questions such as whether models distinguish confounders from true pathological signals, how spurious correlations in datasets affect generated reports, or how causal knowledge can be embedded and evaluated are largely unaddressed. As RRG systems move closer to clinical deployment, understanding and formalising causal reasoning is essential for robustness, generalisation, and patient safety. This gap

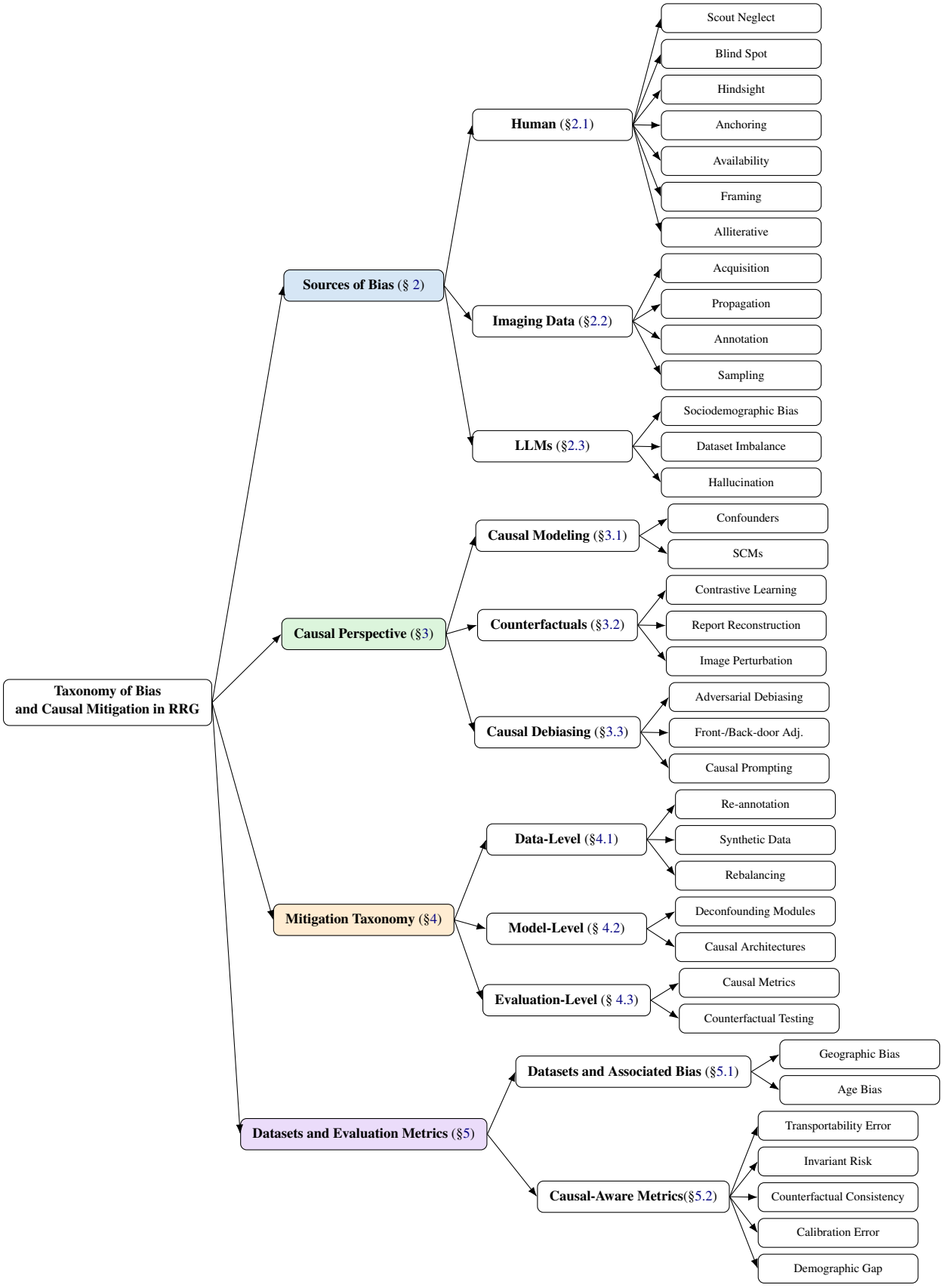

Figure 3: Taxonomy of bias sources and causal mitigation strategies for radiology report generation.

motivates the need for a dedicated survey on causal perspectives in RRG. Such a survey would complement existing methodological and architectural reviews by analysing datasets, models, and evaluation protocols through the lens of causality, and by identifying concrete pathways toward causally grounded and clinically trustworthy report generation systems.

C Prior Work in Medical Domain using Causal Inference

As discussed in the introduction (§ 1), causal reasoning is being used in medical imaging field due to its advantages. In this section, we discuss various prior works in medical domain that used causal inference techniques.

Causal reasoning in medical imaging broadly (Castro et al., 2020), or proposed specific causal methods such as counterfactual augmentation for RRG (Song et al., 2023) . In the study (Kheya et al., 2024; Tripathi et al., 2023) often focuses on either technical statistical bias or social bias in AI models separately. (Tripathi et al., 2023) uniquely highlights how cognitive biases in radiologists, social biases embedded in language models, and statistical confounding in imaging data interplay and propagate unfairness in automated report generation, providing a holistic causal framework to understand and mitigate these intertwined biases.

Beyond acquisition, contextual confounders in the image can spuriously correlate with disease. For instance, an EKG lead or a chest tube in an X-ray might correlate with a diagnosis, but is not a causal feature of the pathology (Castro et al., 2020; Koçak et al., 2025). A model might erroneously rely on such cues. Ideally, clinical findings should be derived from the image itself. However, many models are trained using correlations, which can mistakenly be treated as causal relationships.

Shi and Norgeot (2022) proposes a unified framework that organizes causal inference methods by the level of decision-making they target like individuals, groups, or entire populations. The paper provides existing healthcare applications and shows that current medical studies rely on a narrow set of causal techniques and lag behind other fields and also provides a practical schematic to guide researchers and healthcare stakeholders in selecting appropriate causal methods and interpreting their results.

Martí-Bonmatí (2021) proposes his work proposes a data-centric research framework for medical imaging that critically examines causal inference and its associated uncertainties. Within the proposed data-centric, observational, and causal-inference-based research approach in radiology is positioned as a computational and epidemiological discipline for precision medicine, relying on longitudinal observational designs and case–control analyses. Causal inference is performed on closed, retrospectively collected cohorts, where researchers do not intervene clinically but leverage secondary data to derive consistent causal insights.

Lv et al. (2025) proposes the Causal Intervention Segmentation Network (CiSeg), which integrates causal inference into Unsupervised domain adaptation (UDA) using a Structural Causal Model (SCM) to separate causal factors from bias. A Counterfactual Disentanglement module decomposes latent features into causal and bias components, while prototype-guided contrastive learning and causal-bias residual alignment improve cross-domain consistency.

Xu et al. (2025) propose a causal inference framework for Medical Visual Question Answering (MedVQA) that addresses cross-modal bias by introducing an explicit visual–textual causal graph and a front-door adjustment to mitigate unobserved confounders between images and questions.

Liu et al. (2025) propose Heterogeneous Causal Disease Pattern Detection (HCDPD), a causal inference framework that models how early-stage diseases give rise to latent disease patterns and corresponding organ-level changes visible in medical images. The method is formulated within a potential outcomes framework with multivariate responses, making it suitable for heterogeneous patient populations and relatively homogeneous control groups. Using Bayesian inference, HCDPD estimates both direct and indirect causal effects.

D About Mitigation Ability of Causal Inference for various RRG biases

In Section 3, Table 1 summarizes biases in RRG and the extent to which causal inference mitigates them. In this section, we provide a more detailed analysis of the conditions under which biases can be fully mitigated, partially mitigated, or remain fundamentally resistant to mitigation.

Causal inference provides a range of methodological tools, including backdoor adjustment,

mediation analysis, frontdoor identification, and selection-aware modeling. However, these tools are not universally applicable to all types of bias encountered in RRG. For example, backdoor adjustment is effective when relevant confounders—such as age, sex, or care setting—are observed and can be conditioned on. In contrast, when confounders are unobserved but intermediate variables, such as clinical workflows or reporting conventions, are available, mediation-based or frontdoor approaches are more appropriate. Nevertheless, some biases, particularly selection bias and collider bias that arise from dataset curation and clinical inclusion criteria, remain difficult to address because key variables are unobserved or causal effects are not identifiable from the available data.

Causal inference can correct spurious correlations when the causal effect is identifiable via the backdoor criterion. All common causes of the evidence (e.g., imaging features or AI outputs) and the outcome (diagnosis or decision) are observed. Biases such as framing, availability, and hindsight satisfy these conditions because they arise from contextual information that influences decisions after the true diagnostic signal is available and can therefore be explicitly modeled or removed. By accounting for these contextual factors, causal methods can eliminate spurious associations through counterfactual reasoning.

Causal inference is only partially effective when key assumptions required for identifiability such as the absence of unmeasured confounding, a correctly specified causal structure, or the existence of a valid adjustment set are not fully satisfied. Biases such as scout neglect and acquisition bias satisfies these conditions in medical imaging, where selection mechanisms induce selection bias that cannot be fully blocked by observed variables.

Causal inference fails when bias arises from unobservable or fundamentally non-identifiable processes. This includes biases introduced during data generation or representation learning, such as systematic label biases, historical diagnostic practices, or large-scale imbalances in training corpora. For example, if certain populations are consistently underdiagnosed or underrepresented in the data, the corresponding counterfactuals are never observed and cannot be recovered through causal adjustment. Similarly, LLM-specific behaviors, such as hallucinated content, emerge from internal representation dynamics shaped by opaque pretraining processes. Because the causal relationships between training data, internal representations, and generated outputs are not explicitly observable or interventionally accessible, causal correction at inference time is not feasible.

E Mapping RRG Biases by Clinical Impact and Mitigation Complexity

Figure 2 in Section 3 presents a conceptual mapping of different bias types in RRG systems along two critical dimensions: severity and mitigation difficulty. In this section, we discuss about this conceptual mapping in more detail.

The matrix reveals a stratification of biases based on both clinical risk and tractability. In the upper-right quadrant, biases such as Dataset Bias, Confounding Bias, Hindsight Bias, Anchoring Bias, and Alliterative Bias cluster together, indicating that they are simultaneously high-impact and difficult to mitigate; these biases are typically rooted in data collection practices, latent causal structures, or human cognitive tendencies that propagate through model training and evaluation, making them particularly dangerous for clinical reliability and fairness. Slightly below but still on the high-difficulty side, Visual-Linguistic Misalignment and Blind Spot Bias highlight challenges arising from multimodal representation gaps and unobserved failure modes, which can silently degrade report quality despite strong aggregate metrics. These biases occur when the model fails to attend to subtle but important visual findings or when textual generation does not faithfully reflect image correctly. In contrast, the upper-left quadrant contains Framing Bias, which has high impact but relatively lower mitigation difficulty, suggesting that careful task formulation, prompt design, and reporting standards can substantially reduce its effects. The lower-left quadrant groups Language Bias, Scout Neglect Bias, Availability Bias, and Selection Bias, reflecting biases that tend to have lower downstream clinical impact and are comparatively easier to address through data balancing, lexicon control, and improved sampling strategies. Notably, the lower-right quadrant is sparsely populated, implying that biases which are both low-impact yet hard to mitigate are less prominent or less consequential in RRG. Although they are easier to mitigate through data cleaning, balanced sampling, or controlled vocabularies, ignoring them can still skew model behavior and evaluation. The matrix therefore encourages a tiered mitigation strategy: address low-difficulty biases

early as hygiene factors, while dedicating focused methodological and causal modeling efforts to high-severity, hard-to-mitigate biases that most directly threaten clinical trust and generalization. Overall, the matrix serves as a prioritization tool, emphasizing that the most urgent research attention should focus on causally rooted, high-impact biases that resist surface-level fixes and require principled causal modeling, dataset redesign, or counterfactual evaluation strategies.

F Causal Inference pipeline for RRG

As discussed in Section 3, that biases in RRG can be controlled by causal modeling of RRG pipeline. In this section, we outline the causal inference pipeline in RRG, showing how factors affect image acquisition, representation learning, report generation, and evaluation, leading to distinct biases.

Figure 4 presents a causal inference–oriented pipeline for RRG, explicitly modeling how clinical, demographic, institutional, and temporal factors propagate through image acquisition, representation learning, report generation, and evaluation, while giving rise to distinct classes of bias. At the upstream level, patient demographics (D), clinical history (P), clinical context (C), and unobserved confounders (U) jointly influence imaging workflows, including the choice between scout images and diagnostic images. The figure highlights scout neglect bias as a key early-stage failure mode, where preliminary or low-quality views are systematically underutilized or ignored, leading to distorted image feature representations (I). Importantly, these image-level biases are not purely technical artifacts but are causally downstream of social, institutional, and clinical processes, emphasizing that representation bias emerges before any language generation occurs.

The middle of the pipeline comprises of temporal and cognitive dependencies, particularly through previous reports (R_{prev}) and previous findings (F_{prev}). These historical variables, coupled with training order and exposure history, introduce availability bias and alliterative bias, where models over-rely on recently seen or frequently occurring patterns rather than the true underlying clinical state (F: true findings). This section makes explicit that RRG systems are not memoryless predictors; instead, they encode temporal feedback loops in which prior text influences current predictions, thereby entangling causal signal with learned reporting habits. The presence of model parameters (M) as a central mediator illustrates how architectural choices and learned weights consolidate these biases, creating blind spots that systematically affect certain pathologies, populations, or rare findings. Multiple biases converge at the level of the generated report (R), including framing bias, hindsight bias, and blind spot bias, often amplified by artifacts and spurious correlations (A) learned during training. Crucially, the evaluation loop—via clinical outcomes (O) and automatic metrics (E) such as BLEU, ROUGE, and Clinical F_1 feedback into model development, potentially reinforcing biased behaviors when evaluation criteria are misaligned with clinical correctness. By explicitly separating findings, reports, outcomes, and metrics within a single causal graph. Many failures in RRG arise not from model capacity limitations, but from misidentified causal targets and biased feedback mechanisms. As such, this motivates the need for causal interventions, counterfactual evaluation, and bias-aware training strategies that operate across the entire pipeline rather than at isolated stages.

G From Confounding to Mediation: A Causal View of Bias in RRG

As noted in Section 2 and 3, biases are interrelated with each other and hence, it is very important to study their interrelationships in RRG. In this section, we provide a structured causal view of bias interrelationships

Figure 5 presents a structured causal view of bias interrelationships in RRG, organized around causal mechanisms such as confounding, collider bias, measurement bias, and mediation. Dataset bias, confounding bias, and alliterative bias are shown as primary sources that influence anchoring bias, indicating that systematic properties of the data and reporting conventions shape the initial hypotheses or reference points used by models and evaluators. Availability bias, although shown as less severe, still acts upstream and contributes indirectly by skewing which patterns are most frequently learned or recalled. This structure emphasizes that many apparent reasoning errors are not isolated failures, but are causally inherited from biased data distributions and institutional practices.

The middle layer highlights collider and measurement biases, where interactions between upstream factors distort evaluation and interpretation. Selection bias and hindsight bias appear within the

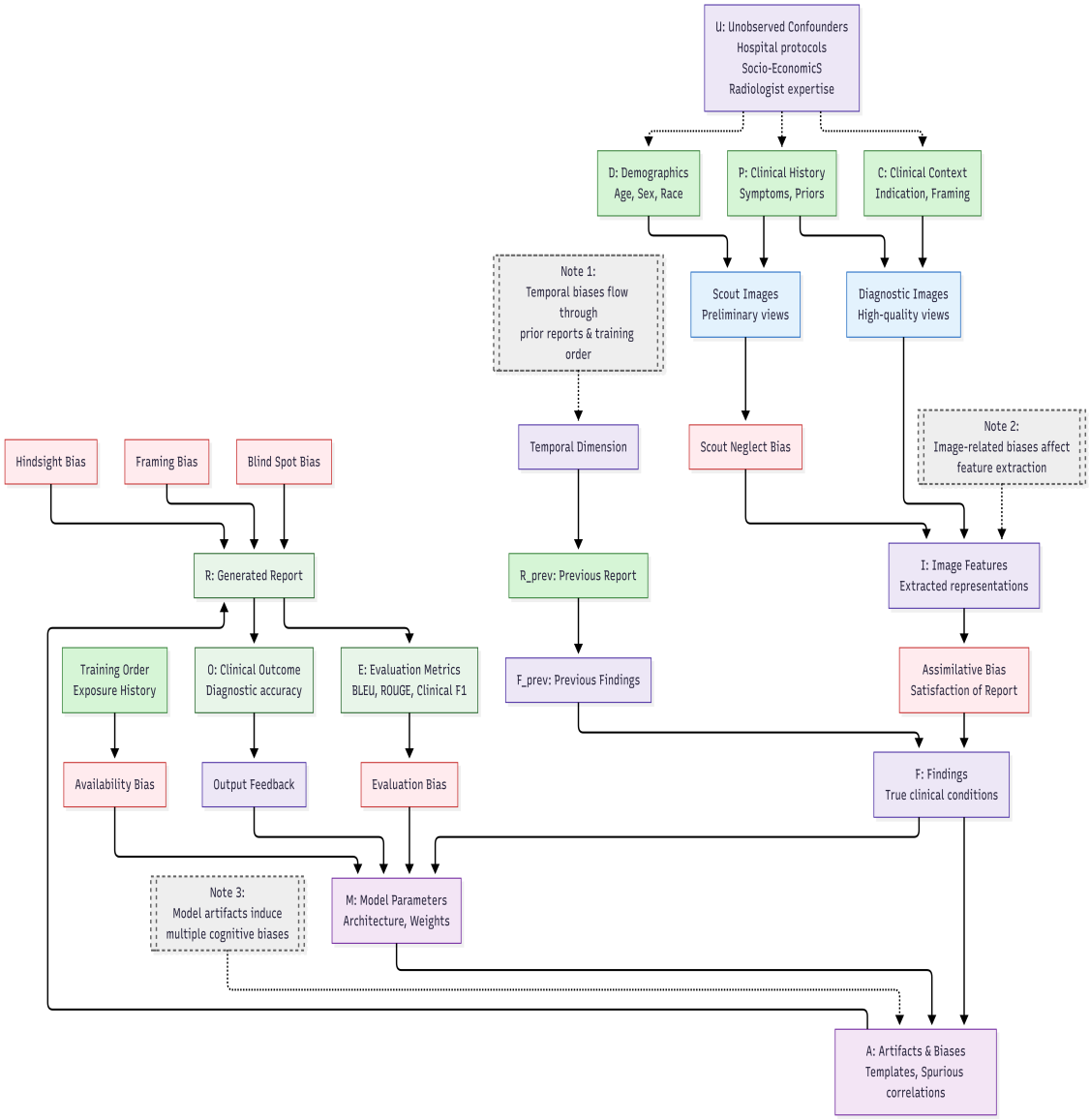

Figure 4: Detailed directed acyclic graph illustrating biases in the RRG pipeline, including unobserved confounders.

collider bias region, indicating that conditioning on observed outcomes or selected samples induces spurious correlations. Hindsight bias directly feeds into evaluation bias, showing how knowledge of ground-truth outcomes contaminates model assessment. In parallel, visual–linguistic misalignment is placed under measurement biases, capturing errors that arise when visual evidence and generated text are not causally aligned. The arrows from anchoring bias and dataset bias to this node suggest that misalignment is not merely a modeling issue, but a consequence of biased supervision and entrenched reporting patterns.

The lower part of the figure focuses on mediation biases, where earlier distortions propagate into final model behavior and clinical interpretation. Framing bias acts as a mediator between upstream biases and downstream language and scout neglect biases, demonstrating how task formulation and report structure shape what the model ultimately expresses. Blind spot bias emerges as a downstream consequence of evaluation bias and hindsight bias, reflecting systematic failure modes that remain invisible under biased assessment protocols. Overall, the diagram conveys a key message: biases in radiology report generation are causally connected rather than independent. Effective mitigation therefore requires intervening at upstream confounders and evaluation practices, rather than addressing surface-level language errors alone.

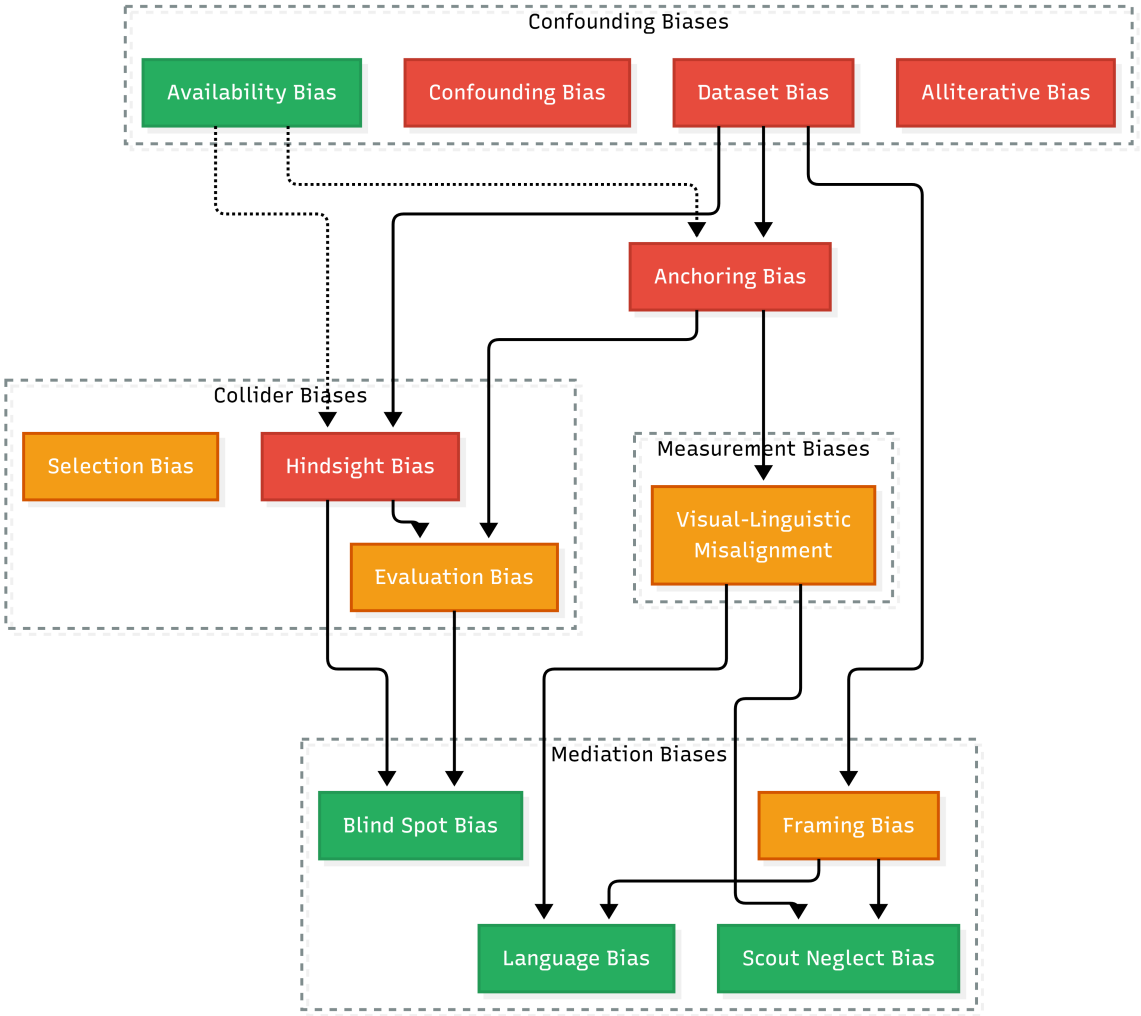

Figure 5: Causal view in RRG from Confounding to Mediation

H Detailed Future Research Directions

In Section 6, we noted that despite recent progress, several concrete and causally grounded research directions remain open for RRG. In this section, we outline specific challenges, methodological gaps, and high-priority opportunities, explicitly linking them to underlying causal assumptions, bias mechanisms, and mitigation strategies.

Unlike classification, RRG produces free-form clinical text, where bias may appear through wording choices, or inclusion of clinically irrelevant risk factors for specific demographic groups. Existing metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) do not capture these disparities. Future work should focus on causal fairness criteria for text generation, for example, by evaluating whether report content is invariant under counterfactual changes to sensitive attributes when pathology is held fixed. Developing counterfactual and group-conditional evaluation metrics aligned with causal estimands is a high-priority direction.

The image encoder and the language generator are treated as a simple pipeline in many RRG models. However, biases often arise because visual features and language priors influence each other during generation. For example, the language model may rely on frequent or stereotypical phrases learned from the training data and overemphasize certain findings when specific visual patterns or contextual cues are present, even if those patterns are not clinically decisive. Modeling the joint causal structure between patient attributes, image appearance, intermediate visual findings, and report text can help identify whether biased wording comes from the visual representation, reporting conventions in the data, or the language model itself. This is particularly important for large multimodal models, where tightly coupled vision–language representations can hide the source of bias and make targeted mitigation more difficult.

The deployment of large multimodal models introduces new causal risks. These models may inherit and amplify biases from both medical and non-medical pretraining data. Future work should investigate causal auditing and intervention strategies for foundation models, including targeted counterfactual testing and controlled fine-tuning, to ensure that performance gains do not come at the cost of fairness or clinical validity.

I Guidance for Radiologists and Clinical Practitioners

Radiologists and clinicians play a key role in identifying clinically meaningful subgroups and real-world failure modes, such as differences in performance for pediatric or elderly patients. Their expertise is essential for defining which disparities matter in practice. At the same time, AI researchers should develop and apply causal methods that can diagnose and mitigate these issues in a principled way.

Prospective clinical trials of RRG systems should explicitly include fairness evaluation, for example by monitoring whether report accuracy and clinical usefulness remain consistent across patient populations. Such evaluations can help ensure that improvements in overall performance do not mask systematic errors affecting specific groups.

On the modeling side, progress in causal representation learning and generative modeling offers opportunities to design more effective data balancing and augmentation strategies that respect underlying causal structure rather than relying on superficial correlations. Ultimately, fairness in radiology report generation is not only a technical concern but also an ethical one, as biased reports can directly influence clinical decisions and patient outcomes.

J Causal Inference in Medical imaging

As discussed in the introduction (§ 1), causal inference is highly useful in medical imaging fields like RRG. In this section, we aim to provide the details of terminologies and concepts involved in causal inference.

`Why do things happen the way they do?` This question is essential in many fields, from healthcare and finance to law. Humans can look for the cause and effect (Penn and Povinelli, 2007), leading to amazing inventions and progress. Often in healthcare, we observe the symptoms or outcomes of a medical condition, but the actual causes remain hidden. For instance, a patient might have a persistent cough and shortness of breath. While these symptoms are visible to the doctor, the underlying cause remains unseen, such as a bacterial infection, long-term smoking, or exposure to air pollutants. Doctors must deduce the hidden causes behind the visible symptoms. In this case, diagnostic tests like chest X-ray scans can help uncover the cause, but some more

factors may be involved which may require other tests for correct diagnosis. The same applies across healthcare, where we often see the effects (symptoms).

To better understand cause and effect, the following two things must be considered (Pearl, 2009)

- **Thinking beyond what we see:** To fully understand an event, we need to think about unseen causes, even if they are not immediately visible. For example, in a chest X-ray, we see the effects of lung disease, but we need to think about the patient's medical history, environmental exposure, or other hidden factors that contributed to the disease.

- **Linking of unseen reasons:** We need to connect the unseen causes to the data we observe. In the case of medical imaging, this means linking of patients medical background, genetics, and lifestyle to the patterns that is detected in their X-rays.

To address these challenges, a mathematical framework has been developed known as a structural causal model (SCM). This model maps out how different causes lead to various effects, helping us to understand not only what is happening but why it is happening. It shows the path from the unseen cause to the observed effect, enabling us to make more informed decisions based on a clearer understanding of the underlying factors.

In the field of medical imaging, SCMs can help us better interpret chest X-rays by linking patterns in the images to the underlying health conditions that caused them. This is especially important in cases where AI models struggle to differentiate between correlations and true causations. By applying causal inference, AI systems can move beyond pattern recognition to provide more meaningful insights into a patient health, offering not just a diagnosis but an explanation of why that diagnosis was made.

J.1 Correlation is not causation

Correlation refers to a statistical relationship between two variables. Correlation does not imply a causal relationship. In other words, just because two variables are correlated does not mean that one variable causes the other to change. **Causation** implies a direct cause-and-effect relationship between two variables. Changes in one variable directly lead to changes in the other variable.

Interpreting radiology images is a crucial aspect of diagnosing conditions. However, it should be noted that while correlations between findings taken from radiology reports and clinical outcomes are often observed, it may be the case that there is no causal relationship between them. Correlations do not necessarily imply causation, due to the following reasons:

- **Confounding variables:** Confounding variables refer to those factors which lead to causally unrelated events. Due to these hidden variables, correlation of observed between these causally unrelated events. Confounding leads to a disagreement between the calculus of conditional probabilities (observation) and do-interventions (actions). Real-world examples of confounding are a common threat to the validity of conclusions drawn from data. For example, in a well known medical study a suspected beneficial effect of hormone replacement therapy in reducing cardiovascular disease disappeared after identifying socioeconomic status as a confounding variable (Humphrey et al., 2002; Barocas et al., 2023).

- **Reverse Causation:** The identified feature might be a consequence of the reported finding, not the cause. Suppose that an AI model had reported presence of nodule in Chest X-Ray images with a corresponding report of lung cancer. This correlation may seem wrong as the nodule could be a benign tumor, not cancerous. A patient has a smoking history in the past that could be a risk factor for lung cancer, which influences both the nodule formation and the development of cancer.

- **Selection bias:** Selection bias is a common concern in medical research and clinical practices such as the interpretation of radiology images. It occurs when the selection of subjects for analysis is not random or true representative of the population. In RRG, selection bias can occur if certain patient demographics, such as age, gender, or pre-existing conditions, disproportionately influence the findings reported (i.e., demographic variables are confounding variables), e.g. doctors want to understand how chest X-ray findings relate to lung cancer. But instead of including people from all age groups and different places, they only look at elderly patients from one specific

area. This might make the results less useful because they only reflect what is happening in that particular group of older people. So, the findings might not apply to younger patients or those from other backgrounds. Similarly, if doctors study how chest X-rays and smoking are connected, but only focus on people who are already sick with lung problems, it seems like smoking is more closely linked to those lung issues than it is. It is because they are not considering healthier people who smoke and are healthy.

J.2 Structural Causal Model

A structural causal model (SCM) $\mathcal{M}$ is a mathematical framework that represents causal relationships between variables. $\mathcal{M}$ is a 4 tuple $(\mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathbf{U}))$ (Pearl, 2009), where

- **External variables or Exogenous variables** are represented by the set of exogenous random variables ($\mathcal{U}$). They are assumed to be independent of each other and have the same probability distribution across all observations.
- **Internal variables or Endogenous variables** are represented by $\mathcal{V} = V_1, V_2, \cdots, V_n$. Their values are determined by other variables in the model, which can include both external factors (from $\mathcal{U}$) and other internal variables (from $\mathcal{V}$ themselves).
- **Causal relationships** is represented by $\mathcal{F}$, which is a set of structural equations $\{f_1, f_2, \cdots, f_n\}$, where each function describes how an internal variable depends on external and other internal variables. $f_i : U_i \cap Pa_i \rightarrow V_i$, where $U_i \subseteq \mathbf{U}$, and $Pa_i \subseteq \mathbf{V} \backslash V_i$ and $\mathcal{F} : \mathbf{U} \rightarrow \mathbf{V}$. Equation 1 captures the causal relationships in the system.

  Each Structural equation f_i is of the form,Pa_i is the set of parent variables of V_i, i.e., the variables directly influencing V_i:

$$\mathbf{V_i} = f_i(Pa_i, \mathbf{U_i}) \tag{1}$$

- $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$

J.3 Confounding

Confounding refers to a scenario where a shared cause, possibly unobserved, masks the causal connection between two or more variables. In causal inference, we can precisely define a causal effect of X on Y as confounded if the probability of Y given X equals x ($p(Y|X = x)$) is not equal to the probability of Y given an intervention on X ($p(Y|\text{do}(X = x))$), indicating that collider bias is a form of confounding. It poses a significant challenge in analyzing observational data. Imagine you want to understand why someone might get a headache. There are many factors involved, like stress, lack of sleep, and even dehydration. Causal DAGs can visually show how these things might be connected. For instance, stress could lead to a lack of sleep, which in turn could cause a headache. There could be other factors that are often ignored. Maybe someone has a headache because they're stressed, but they're also dehydrated because they haven't been drinking enough fluids. Dehydration itself can cause headaches too. This is where do-calculus comes in. It helps us to figure out which factors (like dehydration) are likely to be taken into account how stress truly affects headaches, without any misleading information.

J.4 Front-door and Back-door criterion

The frontdoor criterion is useful when there are hidden factors that influence both the image and the report, and these factors cannot be directly measured. However, if there is an observable intermediate step on the causal path—such as structured clinical findings—then we can still reason about causality. In radiology report generation, findings like cardiomegaly or opacity naturally play this intermediate role.

In RRG, the causal process follows a clear sequence: the image leads to clinical findings, and the findings lead to the final report. Even if unobserved factors affect how reports are written, frontdoor adjustment can recover the causal effect of the image as long as these hidden factors do not directly influence the extracted findings.

Under the frontdoor assumptions, the causal effect is identifiable as given by equation 2

$$\begin{aligned} P(R \mid \text{do}(X = x)) = \sum_m P(M = m \mid X) \\ \sum_{x'} P(Y \mid M = m, X = x') \\ \times P(X = x') \end{aligned} \tag{2}$$

where X is the input image, M is the extracted clinical findings (mediator), Y is the generated, and U is the unobserved confounder.

Many RRG systems work in two steps. First, they predict clinical findings from the image. Then, they generate the report using only those findings. If the findings capture all important medical information in the image, this design follows the frontdoor principle. Using a two-stage approach is especially helpful when information such as prior history or reporting style is missing or unknown. Because the report is generated only from predicted findings, the model is less influenced by hospital-specific language or reporting habits. As a result, the generated reports are more faithful to the image and easier to interpret. It also supports counterfactual reasoning: if patient demographics change but the findings stay the same, the report should remain unchanged.

The backdoor criterion is used when it is affected by both the radiology image and the generated report. These variables are called confounders. If the confounders are ignored, the model may learn patterns that are correlated with disease but not truly caused by the image.

In RRG, common confounders include patient age and sex, scanner type (portable vs fixed), and care setting (ICU vs outpatient). These factors can change how images look and also influence how radiologists write reports. For example, ICU patients often have portable X-rays and more severe conditions, so a model may wrongly link portable scanner artifacts to disease. Backdoor adjustment removes these misleading paths by conditioning on confounders, so the model learns causal relationships rather than correlations. Backdoor criterion is given by equation 3

$$P(Y|\text{do}(X)) = \sum_{z} P(Y|X, Z = z)P(Z = z) \tag{3}$$

where X is the input image, Y is the generated radiology report (or a finding), and Z is the confounders (e.g., age, sex, ICU status)

For a detailed discussion of causality and causal inference, we refer the reader to standard references such as (Barocas et al., 2023; Pearl et al., 2016).

K Information about use of AI Assistants

We used Gemini for minor writing and presentation improvements.