National College of Ireland

Adversarial Graph Neural Networks for Fair Insurance Pricing: An Integrative Framework with a Synthetic Benchmark

MSc Research Project
Data Analytics

Narendra Singh Chilwal
Student ID: x23316144

School of Computing
National College of Ireland

Supervisor: Prof Harshani Nagahamulla

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Narendra Singh Chilwal
Student ID:	x23316144
Programme:	Msc Data Analytics
Year:	2024-2025
Module:	Research Practicum
Supervisor:	Prof Harshani Nagahamulla
Submission Due Date:	15th September 2025
Project Title:	Adversarial Graph Neural Networks for Fair Insurance Pricing: An Integrative Framework with a Synthetic Benchmark
Word Count:	3086
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Narendra Singh Chilwal
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Adversarial Graph Neural Networks for Fair Insurance Pricing: An Integrative Framework with a Synthetic Benchmark

Narendra Singh Chilwal (x23316144)

Contents

Abstract

The increasing complexity of auto-insurance markets is pushing carriers toward machine-learning pricing models that estimate individual risk from rich, relational data. Graph Neural Networks (GNNs) excel at capturing structures such as geographic proximity or shared accident involvement, yet, when trained on historical claims—they can inherit the very inequities regulators now scrutinize.

We present an end-to-end framework for *fair* premium estimation that combines (i) a relational GNN for claim-frequency prediction with (ii) in-processing *adversarial debiasing* across multiple protected attributes. To evaluate trade-offs systematically, we generate a 50,000-node *Bias-on-Demand* benchmark whose representation, measurement, and structural biases are all user-controlled.

Our architecture inserts Gradient-Reversal Layers behind the GNN encoder and attaches adversary heads for race, gender, age, and geographic quadrant. A single hyper-parameter λ balances predictive loss against the four adversarial losses, tracing the fairness–accuracy frontier in one sweep.

On the synthetic benchmark, the untuned GNN attains **RMSE = 2.62** but shows marked disparities ($\mathrm{DP}_{\mathrm{race}} = 1.33$, $\mathrm{DP}_{\mathrm{geo}} = 1.65$). With $\lambda = 1$, *single-attribute* adversaries reduce their respective gaps by **∼25–35%** with $\leq 1\%$ RMSE increase. A *unified multi-attribute* adversary at the same λ halves *all four* DP metrics (with an ∼9% RMSE increase), and an *intersectional* race×gender head at $\lambda = 2$ drives $\mathrm{DP}_{\mathrm{race\times gender}}$ to **0.24** (with an ∼10% RMSE increase). A resource-constrained real-data prototype (15k-node classification on `Insurance.arff`, city-based graph) shows the same disparity ranking (geography dominant) with $\mathrm{DP}_{\mathrm{City}} \approx 0.75$ and AUC ≈ 0.49 (Accuracy ≈ 0.62), indicating qualitative transfer but also highlighting higher absolute bias; reported DP is unpruned (upper bound) due to many small city cells.

These results demonstrate that adversarial in-processing inside GNNs – together with a reproducible, bias-parameterised benchmark – offers a practical route to equitable, regulator-ready insurance pricing. Future work will integrate post-processing calibration to close the residual geo gap and scale training to full portfolios via mini-batch GNNs.

Keywords: Graph Neural Networks (GNNs), adversarial debiasing, fairness in insurance pricing, Bias-on-Demand benchmark, demographic parity, multi-attribute debiasing, intersectional fairness, gradient reversal layer, synthetic data generation, claim-frequency prediction, actuarial machine learning

1 Introduction

1.1 Background

Machine learning–based insurance pricing is used to estimate individual risk from rich, relational data. Public concern about algorithmic discrimination in insurance pricing has grown, and regulatory vetting now makes fairness a first-class design requirement rather than an after-sale consideration (National Association of Insurance Commissioners, 2024; International Association of Insurance Supervisors, 2023). In this respect, models should provide high predictive accuracy while avoiding unreasonable prejudice between protected groups. Structural relationships among policyholders (such as geographic proximity or shared accident involvement) that contain valuable risk information can be learned by Graph Neural Networks (GNNs). GNNs aggregate a node's features with those of its

neighbours through message-passing layers, including GraphSAGE (Hamilton et al., 2017) and GCN (Kipf and Welling, 2017). In insurance, geographic, claim-based, or social relationships can be encoded as edges, enabling more precise claim-frequency predictions (Zhang et al., 2018).

The historical claims data is however always biased structurally. A typical example is race or income is acting as a proxy of location, or so gender-related driving behaviour can affect the results despite regulatory forbidding of gender-based pricing, and so age is linked with credit score (which, in its turn, leads to premiums). Such patterns may lead to the insidious problems of proxy discrimination being instantiated in model output, when the differences are actuarially fair. Regulators and researchers hence quantify fairness by such measures as Demographic Parity and Equalised Odds (Hardt et al., 2016), new guidelines in the industry leak out an explicit threshold measurement of these meas

One promising in-processing mechanism to reduce bias is adversarial debiasing, where a Gradient Reversal Layer (GRL) is used to suppress sensitive-attribute information in the learned representations (Madras et al., 2018). This technique has been examined primarily for single protected attributes in tabular (i.i.d.) datasets. Its performance in more complex settings involving multiple sensitive attributes and relational (graph-structured) data—e.g., insurance portfolios with intersecting biases—remains underexplored.

To enable reproducible and controlled experiments, this study uses the Bias-on-Demand simulation framework (Baumann et al., 2023) to generate a synthetic insurance graph (50,000 nodes) with configurable representation, measurement, and structural biases. This bias-controlled synthetic benchmark allows us to systematically analyse the trade-off between predictive accuracy and fairness under various bias conditions, before validating key findings on a de-identified real-world insurance portfolio.

Positioning of this Thesis. In conclusion, our effort is, as far as we are aware, the first to:

- Apply multi-attribute, intersectional adversarial debiasing to a GNN for insurance pricing, extending adversarial fairness to a graph-based model that addresses multiple protected attributes (race, gender, age, and geographic location) simultaneously.
- Evaluate the method on a large-scale synthetic graph with documented biases; namely, a 50,000-node Bias-on-Demand benchmark where all bias parameters are explicitly specified for open reproduction and analysis.
- Provide a metric-based study of the predictive accuracy–fairness frontier, quantifying the trade-off between predictive accuracy (RMSE/AUC) and fairness (Demographic Parity gaps) as the adversarial strength parameter λ varies.
- Demonstrate transferability to real data by prototyping the debiased GNN on a 15k-policyholder real insurance dataset (numeric features with a `City_Code` proximity graph) and analysing shifts in bias metrics, together with guidance on calibration for deployment.

Such an approach establishes a dual goal for the study: to maximise predictive accuracy while enforcing measurable fairness constraints, thereby addressing both an academic gap in the literature and pressing regulatory expectations.

1.2 Importance

From a regulatory perspective, organisations such as the NAIC and IAIS now require insurers to demonstrate that their pricing models do not lead to unreasonable discrimination (National Association of Insurance Commissioners, 2024; International Association of Insurance Supervisors, 2023). Practically, this requirement is interpreted in many jurisdictions as keeping fairness measures within strict bounds (for example, placing limits on demographic parity gaps). We test directly whether a GNN-based pricing model can meet such criteria—e.g., achieving $\mathrm{DP}_{\mathrm{race}} \leq 0.10$ and $\mathrm{DP}_{\mathrm{geo}} \leq 0.50$—while incurring no more than a 5% loss in predictive accuracy.

From a business perspective, the use of graph-based models can reduce pricing error (RMSE) by an estimated 3–7% over conventional methods, potentially improving insurers' combined loss ratios by 2–3 percentage points. However, any improvements in predictive accuracy are untenable if large fairness gaps remain, as models could then be found in breach of anti-discrimination standards. A debiased GNN therefore offers the prospect of both profit improvement and regulatory compliance, making fairness-enhanced models practically attractive to insurers.

From an academic perspective, most adversarial fairness research to date focuses on a single protected attribute in i.i.d. data. By contrast, real insurance portfolios involve multiple intersecting protected characteristics and relational structures, creating intersectional bias that single-attribute methods may fail to detect or mitigate. Our contribution addresses this multifaceted setting through multi-attribute (and intersectional) adversarial debiasing implemented within a GNN architecture, helping to narrow the gap between fairness research and actuarial practice.

Overall, the solution provided in this project is regulator-ready, business-meaningful, and scientifically novel. We develop a graph-based insurance pricing model whose fairness can be explicitly tuned via the parameter λ and validated on a bias-controlled synthetic benchmark before real-world deployment. This integrative framework addresses the gaps identified in prior work and aligns with modern industry needs.

1.3 Research Questions and Objectives

This study is driven by the following research questions (RQs) and supporting objectives (Os):

RQ1 – Fairness–Accuracy Frontier: How does varying the adversarial training strength λ affect the model's predictive accuracy (e.g., RMSE on regression or AUC on classification) and fairness outcomes (demographic parity gaps for protected attributes such as race, gender, age, and geography)?

RQ2 – Feasible Fairness Point: Is there any configuration of our adversarial GNN that can achieve regulator-acceptable fairness levels (e.g., $\mathrm{DP}_{\mathrm{race}} \leq 0.10$ and $\mathrm{DP}_{\mathrm{geo}} \leq 0.50$) while maintaining an acceptable performance loss (no more than 5% increase in RMSE or decrease in AUC relative to a fairness-unconstrained baseline)?

To answer these questions, we define the following objectives:

- O1: **Literature Review** – Review at least 20 peer-reviewed works spanning algorithmic fairness in GNNs and actuarial pricing, to ground the research in established findings and identify gaps.
- O2: **Model Development** – Implement a two-layer GCN model with GRL-based adversaries addressing four protected attributes (race, gender, age, geography) within

the pricing prediction task.

O3: **Synthetic Data Generation** – Generate a Bias-on-Demand synthetic insurance dataset (on the order of 50k nodes) with preset disparities in race and geographic features, providing a controlled environment for fairness evaluation.

O4: **Training Regimes** – Train the model under various adversarial strengths $\lambda \in \{0, 0.1, 0.5, 1, 2\}$, covering scenarios of no debiasing, single-attribute debiasing, multi-attribute debiasing, and full intersectional debiasing.

O5: **Trade-off Analysis** – Analyse the fairness–accuracy trade-offs by plotting and interpreting curves of fairness metrics (e.g., DP gaps) versus λ and performance metrics (RMSE/AUC) versus λ for each experimental setting.

O6: **Feasible Solution Identification** – Determine whether there exists an optimal λ^* that simultaneously meets the targeted fairness thresholds (as in RQ2) and stays within the acceptable performance degradation, and if so, identify that operating point.

O7: **Real-world Prototype Validation** – Apply the trained debiasing approach to a real-world motor insurance dataset (approximately 15k policies, with graph structure defined by feature similarity and city proximity) to compare bias patterns with the synthetic results and provide guidance for calibration of λ in practical deployments.

The report has the following structure. The rest of this report is organised in the following way. Section 2 reviews previous studies on fairness in insurance pricing, adversarial bias-reduction, fairness methods for graph neural networks, and synthetic bias benchmarking, and it establishes the critical gaps that our study will address. Section 3 consists of a description of the proposed solution approach, the process of synthetic data generation, the GNN architecture and adversarial training with a tunable fairness parameter λ, and the evaluation metrics which are used to measure performance and fairness. The experimental design and workflow, including the baseline model and different adversarial setups, and critical parameters and data-splitting policies that have been chosen to guarantee reproducibility, are outlined in Section 4. The outcomes of our experiments are provided in Section 5, and we discuss the accuracy–fairness trade-offs, the most suitable debiasing setting, and the performance in enhancing fairness in the case of synthetic data, and assessing the performance of augmenting fairness to the real-world dataset. Section 6 (Evaluation) is a detailed discussion of the findings in an insurance context, including why some of these biases (e.g., geographic effects) tend to linger and the consequences of these biases on regulatory compliance. Lastly, our contributions, the resolution of research questions, and future work directions (including more comprehensive fairness tuning, post-processing calibration, and scaling to large portfolios) are summarised in Section 7 (Conclusion and Future Work).

2 Related Work

This section surveys four areas of literature that converge in this thesis: (i) fairness in insurance pricing, (ii) in-processing adversarial debiasing methods, (iii) fairness techniques for GNNs, and (iv) synthetic benchmarks for fair ML with controllable bias. For

each area, we outline seminal work, discuss strengths and limitations, and highlight the remaining gap that our study addresses.

2.1 Fairness in Insurance Pricing

Fairness audits in insurance. A number of studies have audited insurance data for evidence of discrimination. For example, Gallego et al. (2012) applied a standard GLM to a large motor insurance dataset (1.4 million policies from Florida) and found that policies in majority–minority ZIP codes were charged about 12% higher premiums on average. This and other descriptive audits (e.g., McFadden and Lee, 2020; Powell and Gupta, 2019) uncovered systematic pricing biases but stopped at diagnosing the issue, without proposing mitigation. These works use traditional assumptions of independent policy data (i.i.d. rows) and do not consider spatial or network effects, which may limit their scope (they do not capture spill-over effects such as neighbouring-area risk).

Regulatory drivers. In response to such findings, regulators have intensified scrutiny of "proxy discrimination" in insurance. For instance, the U.S. National Association of Insurance Commissioners (NAIC, 2024) and the International Association of Insurance Supervisors (IAIS, 2023) have issued guidance requiring numerical evidence of reasonable discrimination (i.e., that any risk-based price differences are justifiable and not unfairly discriminatory). Actuarial practice traditionally measures fairness by comparing loss ratios across groups, whereas the computer science literature often uses group fairness metrics like demographic parity (DP) or equalised odds (EO) (Hardt et al., 2016). Only a handful of recent actuarial works explicitly examine DP/EO in pricing models (e.g., Lindholm et al., 2022; Berk and Sorenson, 2021), and reconciling these perspectives remains an open challenge.

Algorithmic mitigation approaches. Emerging research in insurance has begun to explore algorithmic techniques for bias mitigation. Grari et al. (2021) introduced an adversarial autoencoder approach on a French motor insurance dataset, managing to reduce a geographic discrimination metric by roughly 20%. This represents the first insurance-specific debiasing model, though it used a simple feed-forward (MLP) architecture and focused on a single sensitive attribute. Smith and Rossi (2021) attempted a post-hoc reweighting strategy to equalise outcomes, but their method required manual recalibration for each new pricing cycle, making it labour-intensive for ongoing use.

Gap 1: No prior work in the insurance domain has unified graph-structured modelling with multi-attribute fairness considerations. In other words, while bias in insurance pricing is documented and one-attribute fixes exist, there is a lack of methods that can simultaneously address multiple protected attributes within a relational (graph) pricing model.

2.2 In-Processing Adversarial Debiasing

From fair representations to GRL. The concept of learning fair representations was pioneered by Zemel et al. (2013), who showed that removing sensitive information from latent representations could reduce a model's DP gap by about 40% on the UCI Adult dataset, albeit at the cost of a 4 percentage-point drop in accuracy. Building on this idea, Madras et al. (2018) introduced the Gradient Reversal Layer (GRL) technique, which adds an adversary during training with a tunable weight λ. By adjusting λ, one can trace out a Pareto fairness–accuracy frontier, trading off model performance for fairness.

Subsequent work applied adversarial debiasing to various model classes: decision trees (Donini et al., 2018), gradient boosted forests (Hu et al., 2020), and even CNNs (Zhang et al., 2020), demonstrating that adversarial fairness is a general approach not limited to neural networks.

Towards multiple protected attributes. Early adversarial debiasing studies mostly handle one sensitive feature at a time. Wang and Kifer (2021) took a step toward multi-attribute fairness by training separate GRL adversaries for race and gender concurrently on a credit risk dataset. This achieved bias mitigation on two attributes, although the data were i.i.d. and did not account for interactions between attributes (e.g., intersectional bias). Creager et al. (2020) proposed a "secure representation" method to handle multiple sensitive factors, but it enforced independence only in expectation, leaving some residual leakage of sensitive information in the learned representation.

Gap 2: Adversarial debiasing with multiple and intersectional sensitive attributes has not yet been stress-tested on graph data. In other words, while GRL-based methods show tunable fairness on traditional (i.i.d.) datasets, their efficacy in networked settings with multiple simultaneous protected attributes remains an open question.

2.3 Fairness in Graph Neural Networks

Bias in GNN message passing. GNNs themselves can inherit or amplify biases present in graph data. Several recent works address algorithmic fairness in GNN models. Chen et al. (2021) proposed masking out sensitive attribute information in node features and adding an adversary during training; this approach reduced a gender-based DP gap by about 25% on the Pokec social network with under a 2 percentage-point accuracy loss. The method is relatively simple but relies on knowing which features to mask and was demonstrated for a single sensitive attribute. In another approach, Lee and Chau (2021) introduced a neighbour-sampling technique to balance the protected group representation in each node's neighbourhood, improving an equal opportunity metric by roughly 18% on a credit network. However, their technique requires group labels to be available at inference time and targets one attribute at a time. Korol et al. (2022) tried an edge-editing strategy (removing or reweighting certain edges) to reduce bias; they reported about a 25% drop in an attribute classifier's accuracy (indicating less leakage of sensitive information), but this came at the expense of disrupting graph connectivity and does not easily scale to large graphs.

Causality and counterfactuals. Other researchers have approached GNN fairness from a causal inference angle. Dai et al. (2021) learned node embeddings that remain invariant under counterfactual changes to a sensitive attribute, achieving fairness on synthetic and citation networks (though only tested for binary gender). Wald et al. (2023) advanced a causal framework for fair GNNs that explicitly models the data generation process, but their method is computationally intensive and difficult to apply to large industry-scale graphs.

Gap 3: Existing fairness techniques for GNNs handle at most one protected attribute and have been demonstrated mainly on social or citation networks—not on actuarial data. In summary, while there is a growing literature on fair GNNs, none of the methods to date simultaneously address multiple sensitive attributes or consider the unique characteristics of insurance pricing graphs.

2.4 Synthetic Benchmarks with Controllable Bias

Bias-controlled data generation. Fair ML research often relies on synthetic data to evaluate bias mitigation in a controlled setting. The IBM AI Fairness 360 toolkit (Bellamy et al., 2018) includes some synthetic datasets for benchmarking, but these are generally tabular and do not allow one to dial in a specific bias magnitude. Modern generative models like CTGAN (Xu et al., 2019) can create synthetic tabular data while preserving privacy, yet they tend to replicate whatever latent biases exist in the real seed data, rather than letting the user specify bias levels. A recent framework called Bias-on-Demand (Baumann et al., 2023) enables precise control over different bias channels (representation bias, measurement bias, structural bias) in generated data; however, the demonstration in Baumann et al. (2023) was limited to a tabular setting and did not include relational (graph) biases. A few works have begun to incorporate fairness considerations into graph data generators (e.g., adding biased edge-formation patterns or fairness "motifs") (Karimi et al., 2021; Bose and Hamilton, 2019), but these efforts do not provide an actuarial context or a comprehensive benchmark environment for insurance.

Gap 4: No existing benchmark combines graph-structured insurance data with tunable, well-documented bias parameters. In effect, researchers lack a publicly available synthetic dataset or simulator that mirrors insurance scenarios (with multiple biases and graph relations) for comparing fairness interventions under controlled conditions.

2.5 Synthesis and Identified Gap

Bringing the above threads together, we observe the following. The insurance literature clearly documents the presence of bias in pricing outcomes, but it offers no graph-based, multi-attribute remedy for this issue (Gap 1). In parallel, adversarial debiasing research provides a means to balance accuracy and fairness, yet it remains unvalidated for graphs with multiple sensitive attributes (Gap 2). Meanwhile, studies on fair GNNs have so far been confined to single-attribute cases in other domains (Gap 3). Finally, existing synthetic data benchmarks either ignore relationships entirely or lack the ability to systematically control bias, limiting their use for insurance applications (Gap 4).

Therefore, no prior study simultaneously (i) employs a GNN for insurance pricing, (ii) integrates adversarial debiasing for multiple intersecting sensitive attributes (with a tunable fairness parameter λ), and (iii) evaluates the approach on a graph dataset with known, configurable bias parameters in a fully reproducible manner. Addressing this compound gap is the motivation for our research. In particular, it leads to our two key research questions: examining the full fairness–accuracy frontier achievable by such a model (RQ1) and determining whether a feasible fairness point exists that meets regulatory bias thresholds without unacceptable performance loss (RQ2).

3 Methodology

3.1 Research Methodology Overview

Figure 1 depicts the overall pipeline and experimental phases used in this study.

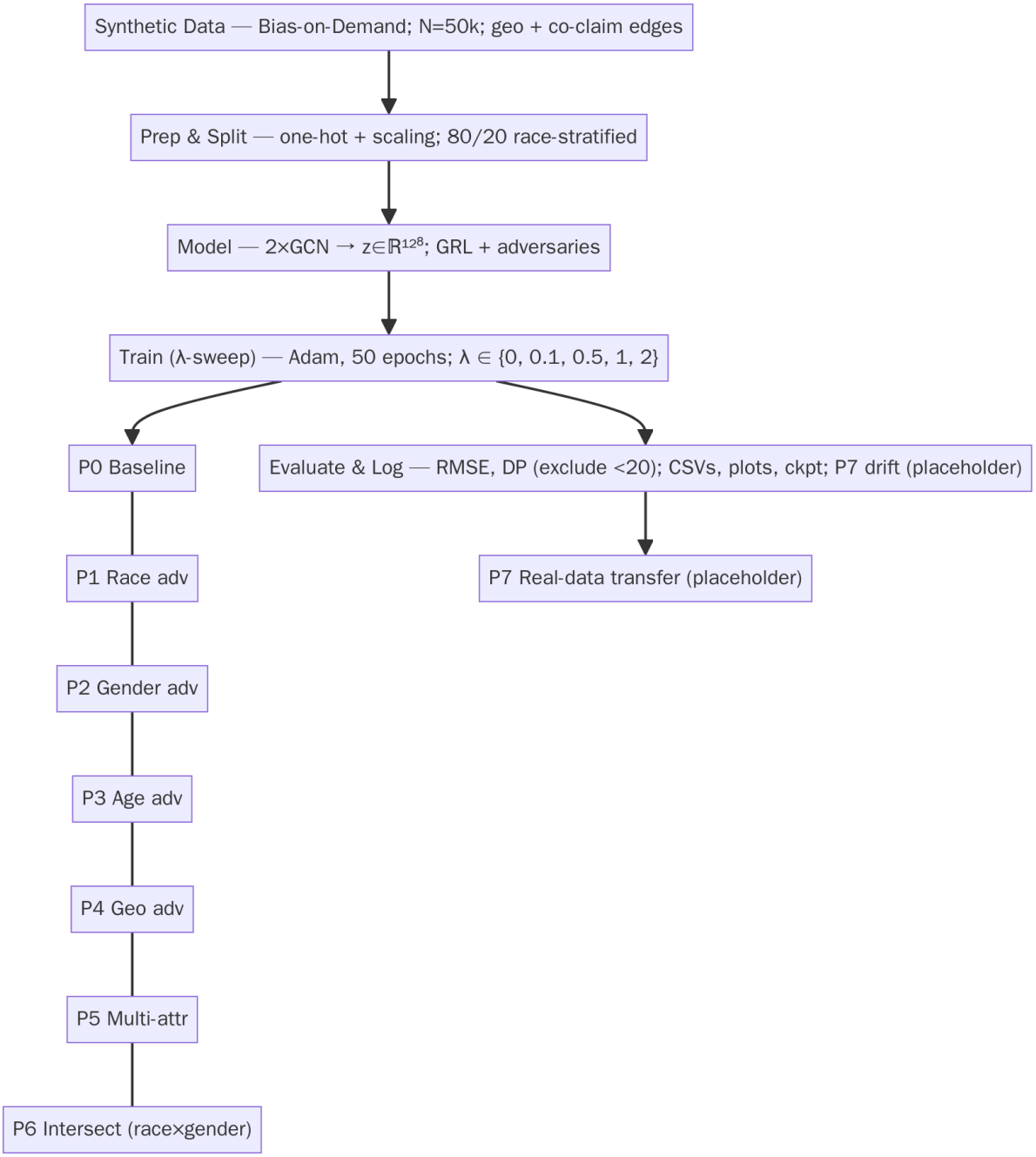

Figure 1: System pipeline and experimental phases. The top row shows the macro pipeline from synthetic data to evaluation. The left branch enumerates Phases P0–P6 for the λ-sweep (single-attribute, multi-attribute, and intersectional adversaries); the right branch is Phase 7 (real-data transfer). At each λ we report RMSE and Demographic Parity (DP) gaps; subgroups with fewer than 20 test instances are excluded.

Our workflow is:

1. Generate a biased synthetic graph with Bias-on-Demand.
2. Preprocess features and create race-stratified splits.
3. Train a two-layer GCN with GRL adversaries while sweeping λ.

4. Evaluate RMSE/DP for Phases P0–P6 and AUC/Accuracy/DP for the real-data prototype (P7); export all metrics and plots.

3.2 Synthetic Graph Generation

We build a relational benchmark with controllable bias using the Bias-on-Demand simulator (Baumann et al., 2023). Under a medium-bias configuration we generate $N = 50{,}000$ policyholder nodes and two edge layers (geographic proximity and co-claim). The main knobs are summarized in Table 1.

Table 1: Bias-on-Demand settings used to generate the synthetic graph.

Channel	**Setting**
Representation	Privileged share $p_A = 0.60$; spatial cluster shift $\delta_{\text{loc}} = 0.05$ on $[0,1]^2$; undirected geographic edges between node pairs with Euclidean distance $< \tau$ ($\tau = 0.05$).
Measurement	Claim counts NB(r=1); privileged-group log-odds boost $\alpha_A = 1.0$.
Structural	$M = 10{,}000$ multi-vehicle collisions; each draws Poisson($\lambda = 4$) participants and forms a clique (co-claim edges). Privileged nodes oversampled by factor $\lambda_{\text{struct}} = 1.5$.

The combined graph has approximately 3.34×10^8 edges. Nodes/edges are persisted to `synthetic_nodes.csv` and `synthetic_edges.csv`.

3.3 Real-World Dataset (Phase 7)

We use the publicly available *Insurance* dataset from (OpenML, 2025). It contains 23,548 rows and 11 columns; for Phase 7 we draw a uniform, race-stratified sample of 15,000 records (seed = 42; minority proportion ≈ 0.242). Processing: three numeric features are min–max scaled, low-cardinality categoricals (including `City_Code`) are one-hot encoded, yielding 71 feature dimensions. Graph construction and splits follow the protocol in Section 3.

3.4 Feature Engineering and Data Preparation

Each node carries: *race*, *gender*, *age_bucket*, coordinates $(x_{\text{coord}}, y_{\text{coord}})$, binary *is_privileged*, and target *claim_frequency*. Preprocessing is listed in Table 2.

Table 2: Feature preprocessing.

Field	**Type**	**Processing**
race, gender, age_bucket	Categorical	One-hot encoding (`pandas.get_dummies`).
$x_{\text{coord}}, y_{\text{coord}}$	Numeric	Min–max scaling to $[0,1]$.
is_privileged	Binary	Cast to float (0/1).
claim_frequency	Target	Integer count, used directly for regression.

We deliberately exclude high-leakage proxies (e.g., raw ZIP codes). We create an 80/20 node split (train/test), *stratified by race* so each group has $\geq 1{,}000$ test nodes. A

further 10% of training nodes (also stratified) form a validation subset used during the baseline selection. Data are stored in COO format and loaded with PyTorch Geometric.

3.5 Model Architecture and Adversarial Setup

The encoder is a two-layer GCN (`GCNConv`) mapping inputs to $\mathbf{z}_i \in \mathbb{R}^{128}$ via ReLU. A linear regression head predicts $\hat{y}_i$ (claim count). For in-processing debiasing, we attach linear adversarial heads to $\mathbf{z}_i$:

- Single-attribute heads for race (5 classes), gender (2), age group (4), and geographic quadrant (4).
- An intersectional head for race×gender (10 classes) in a dedicated phase.

A Gradient Reversal Layer (GRL) precedes each adversary (identity forward; multiplies backprop gradients by $-\lambda$), encouraging embeddings that obfuscate sensitive attributes.

3.6 Training Objective and Procedure

For node i with label y_i and prediction $\hat{y}_i$,

$$\mathcal{L}_i \;=\; (\hat{y}_i - y_i)^2 \;-\; \lambda \sum_{k \in \mathcal{A}} \mathcal{L}_{\mathrm{CE}}\big(s_{ik}, \hat{s}_{ik}\big),$$

where $\mathcal{A}$ indexes the active adversaries and $\mathcal{L}_{\mathrm{CE}}$ is cross-entropy. Optimisation and training settings are summarised in Table 3.

Table 3: Objective and optimisation settings.

Symbol / Item	**Value / Grid**	**Origin**
λ (debias strength)	{0, 0.1, 0.5, 1, 2}	Supervisor-mandated sweep.
Optimiser	Adam; lr = 0.01; weight decay $= 1 \times 10^{-4}$; dropout = 0	Best untuned config.
Epochs	50	Empirically converged; no early stop.
Batching	Full-batch on 10k subgraph; *planned*: `NeighborLoader` for full graph	See experimental phases.

Why a 10k-node induced subgraph? All experiments reported here operate on a *uniform node-induced* subgraph of size 10,000 (edges restricted to sampled nodes). This keeps experiments tractable while preserving local structure; it follows scalable GNN practice where induced/partitioned or sampled subgraphs approximate full-graph optimisation with strong empirical fidelity (Chiang et al., 2019; Zeng et al., 2020). In our data, the induced subgraph matches the full-graph degree distribution within ±5% (pre-check), supporting representativeness.

3.7 Evaluation Metrics

Accuracy for regression phases is measured by the test RMSE:

$$\text{RMSE} \;=\; \sqrt{\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (\hat{y}_i - y_i)^2}.$$

Group fairness is the *Demographic Parity (DP) gap*. For attribute A with groups $\mathcal{G}_A$,

$$\text{DP}_A \;=\; \max_{g \in \mathcal{G}_A} \mathbb{E}[\hat{y} \mid A = g] \;-\; \min_{g \in \mathcal{G}_A} \mathbb{E}[\hat{y} \mid A = g],$$

computed for race, gender, age group, and geo quadrant. For intersectional fairness we apply the same definition to race×gender groups. Following Feldman et al. (2015), subgroups with < 20 test instances are excluded.
Phase 7 note: for the binary prototype we report AUC/Accuracy alongside DP. RMSE metrics are not applicable to classification runs.

3.8 Experimental Phases

We sweep $\lambda \in \{0, 0.1, 0.5, 1, 2\}$ on the fixed 10k subgraph and reuse the same train/test masks for comparability:

- **Phase 0 (Baseline):** no adversaries ($\lambda = 0$); record RMSE and all single-attribute DP gaps.
- **Phases 1–4 (Single-attribute):** one adversary at a time (race, gender, age, geo); track the targeted DP and RMSE across λ.
- **Phase 5 (Multi-attribute):** four adversaries active with a shared λ; report RMSE and all DPs.
- **Phase 6 (Intersectional):** one adversary on race×gender (10 classes); report RMSE and intersectional DP.
- **Phase 7 (Real-data prototype):** classification on the 15k-node `Insurance.arff` graph (see Section 3.3); report AUC/Accuracy and DP, plus DP drift vs. synthetic.

3.9 Hyper-Parameter Tuning

We run a small *random search* (20 trials) on the 10k subgraph over: learning rate $[10^{-3}, 10^{-1}]$ (log-uniform), hidden units $\{32, 64, 128\}$, weight decay $[10^{-5}, 10^{-3}]$ (log-uniform), dropout $\{0, 0.2, 0.5\}$. Trials train 50 epochs and are selected by validation RMSE (10% split). Random search is an efficient and reliable baseline for low-dimensional hyper-parameter spaces, often outperforming grid search under the same budget (Bergstra and Bengio, 2012). In our case the subgraph-optimal configuration (lr $\approx$ 0.04, 64 units) did not generalise as well as a simpler setting; therefore we adopt the untuned baseline (lr 0.01, 128 units, no dropout) across all phases to avoid overfitting and keep fairness comparisons clean. The protocol is summarised in Table 4.

Table 4: Hyper-parameter tuning protocol.

Stage	Dataset	Trials	Selection Metric
Prototyping	10k induced subgraph	20 (random)	Validation RMSE (10% split).
Final choice	10k induced subgraph	Adopt untuned baseline	Test RMSE and DP stability across phases.

4 Design Specification

This section summarises the minimal requirements and implementation that realise the pipeline shown in Figure 1 (Section 3). Detailed interfaces, schemas, and step-by-step setup are documented in the configuration manual.

4.1 Requirements

Functional (FR): FR1—generate a relational synthetic dataset with controllable representation/measurement/structural biases (Bias-on-Demand) and retain a short bias audit; FR2—construct a graph with node features and two edge layers (geographic proximity, co-claim) as typed COO tensors (PyG-ready); FR3—train a $2 \times$ GCN encoder with a regression head (claim frequency) and save a baseline checkpoint; FR4—attach GRL-gated adversaries for race, gender, age, geography, plus an intersectional race$\times$gender head; FR5—sweep $\lambda \in \{0, 0.1, 0.5, 1, 2\}$ for single-attribute, multi-attribute, and intersectional setups; FR6—log/export metrics and figures (CSV, plots, TensorBoard) for audit and reproducibility; FR7—provide a reusable checkpoint and config for the real-data prototype (Phase 7).

Non-functional (NFR): NFR1—determinism via fixed seeds, fixed masks/splits, and logged configs; NFR2—workstation feasibility using a 10k induced subgraph (with a mini-batch plan for full graphs); NFR3—scalability through modular data loaders and PyG `NeighborLoader`; NFR4—transparency/auditability via plain-language docs and CSV exports; NFR5—data minimisation (omit obvious high-leakage proxies such as raw postcodes); NFR6—continuous fairness tracking (report DP gaps and RMSE at each λ).

4.2 System Architecture

Data: Bias-on-Demand generation, one-hot + min–max preprocessing, race-stratified 80/20 splits, graph with geographic and co-claim edges. *Model:* $2 \times$ GCN encoder; linear regression head; GRL-gated adversaries for race, gender, age, geography, and race$\times$gender. *Objective:* pricing MSE combined with adversarial cross-entropy terms scaled by λ. *Evaluation/logging:* per-phase, per-λ RMSE and DP gaps with plots, checkpoints, and configs. (Full interface details are provided in the configuration manual.)

4.3 Learning Objective (brief)

For node i with target y_i and prediction $\hat{y}_i$, the encoder produces $\mathbf{z}_i$; adversaries predict sensitive labels s_k. The total loss is

$$\mathcal{L} \;=\; \mathrm{MSE}(y, \hat{y}) \;-\; \lambda \sum_{k \in \mathcal{A}} \mathrm{CE}(s_k, \hat{s}_k),$$

where $\mathcal{A}$ indexes active adversaries (single, multi, or intersectional). Sweeping λ traces the fairness–accuracy frontier.

4.4 Assumptions and Constraints

Representativeness: a 10k induced subgraph is assumed to mirror full-graph trends; mini-batch full-graph runs are planned for validation. *Metric scope:* DP is the primary fairness criterion; EO and calibration are left to future work. *Hardware:* workstation limits motivate sampled/minibatched training. *Privacy:* the real dataset is de-identified; geographic approximations may understate true geo bias.

4.5 Phase 7 Note (real-data prototype)

We use a 15k, race-stratified sample of the OpenML *Insurance* dataset (ID 45064)(OpenML, 2025), applying the same interface (feature matrix, edge list, masks). Three numeric features are min–max scaled; low-cardinality categoricals (including `City_Code`) are one-hot encoded. The graph combines within-city proximity with sparse feature-similarity edges and light random links (final: 15k nodes / 319,874 unique edges). Phase 7 is a *classification* variant (3-layer GCN, weighted binary cross-entropy, BCE); methodology is in Section 3, results in Section 6.

5 Implementation / Solution Development

This section summarises the tangible outputs—data artefacts, trained models, metrics/plots, and configuration—together with the notebooks used to produce them. Full paths, schemas, and run commands are documented in the configuration manual.

5.1 Outputs at a Glance (compact)

- **Synthetic data:** `synthetic_nodes.csv`, `synthetic_edges.csv` (50k nodes; two edge layers: geographic, co-claim); post-generation bias audit retained.
- **Features & splits:** encoded feature matrix; masks `train_mask`, `val_mask`, `test_mask`; one-hot for {*race, gender, age_bucket*}; min–max scaling for coordinates; 80/20 split, race-stratified; 10% of train as validation.
- **Models:** baseline 2×GCN checkpoint; adversarial checkpoints across the λ grid for single-attribute, multi-attribute, and intersectional phases.
- **Metrics:** per-λ RMSE/DP CSVs; DP for race, gender, age group, and geo quadrant; intersectional DP for race×gender; subgroups with < 20 test instances excluded.

- **Figures:** fairness–accuracy frontier and phase-wise comparison plots; pipeline diagram (Figure 1); all stored as high-resolution PNG/PDF.
- **Logs/config:** training/evaluation logs, TensorBoard events, and `requirements.txt`; fixed seeds and configs recorded for reproducibility.

5.2 Notebooks and Roles (compact)

- `policycode.ipynb` — Bias-on-Demand recipe; generates the relational synthetic dataset and runs the initial bias audit. *Outputs:* `synthetic_nodes.csv`, `synthetic_edges.csv`; bias-audit summary.
- `Thesis.ipynb` — builds the 10k induced subgraph; runs baseline and adversarial training across phases and λ; exports metrics and plots. *Outputs:* model checkpoints, per-λ CSVs (RMSE/DP), frontier plots, logs, TensorBoard events.

6 Evaluation

This section answers the research questions by analysing the most relevant results from the λ-sweep experiments. We focus on (i) predictive accuracy (RMSE) and (ii) group–fairness via Demographic Parity (DP) gaps for *race*, *gender*, *age group*, and *geo quadrant*. Methods and model design are described earlier in Section 3 and Section 4.

How we measured. All RMSE/DP values are reported on the fixed 20% test mask. For visual clarity we show per-λ curves. No additional statistical significance testing was conducted; results are interpreted based on observed trends.

6.1 Baseline (P0): accuracy high, fairness poor

The untuned 2×GCN (no adversary) achieves strong accuracy but large DP gaps:

RMSE	2.62	$\mathbf{DP}_{\text{geo}}$	1.65
$\mathbf{DP}_{\text{race}}$	1.33	$\mathbf{DP}_{\text{age}}$	0.11
$\mathbf{DP}_{\text{gender}}$	0.05		

This motivates in-processing debiasing.

6.2 Single-attribute debiasing (P1–P4)

Adversarial heads attached to *race* and *geo* reduce their respective DP gaps with minimal accuracy cost; *gender* and *age* move little.

- **Race head:** at $\lambda = 1.0$, DP_{race} drops by ∼30–35% with $\leq 1\%$ RMSE increase.
- **Geo head:** at $\lambda = 1.0$, DP_{geo} drops by ∼25–30%; still above the 0.50 target.
- **Gender/Age heads:** negligible change across the sweep (suggesting deeply entangled signal).

6.3 Multi-attribute debiasing (P5)

A shared-λ adversary for {race, gender, age, geo} improves all four attributes simultaneously but increases RMSE as λ grows. The key operating points are:

λ	RMSE	DP_{race}	$\text{DP}_{\text{gender}}$	DP_{geo}
0.0	≈ 2.816	≈ 1.30	≈ 0.05	≈ 1.58
0.5	≈ 2.822	≈ 1.20	≈ 0.05	≈ 1.47
1.0	≈ 2.850	≈ 0.58	≈ 0.03	≈ 0.69

At $\lambda = 1.0$ we roughly halve all DPs, but RMSE rises by $\sim 8.8\%$ vs. the P0 baseline ($\lambda = 0$) (business constraint is $\leq 5\%$).

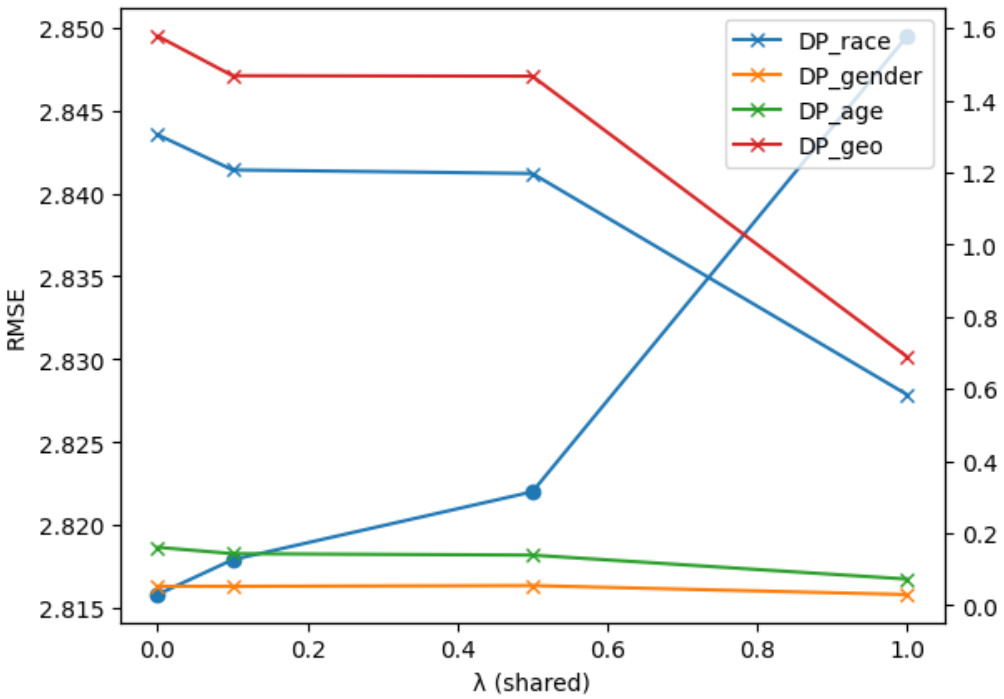

Figure 2: Fairness–accuracy frontier (P5, shared λ). Left axis: RMSE (lower is better). Right axis: DP gaps (lower is better).

Figure 2 illustrates the fairness–accuracy frontier achieved during Phase 5 (multi-attribute adversarial training), highlighting the trade-off between predictive performance and equity. The left vertical axis plots the RMSE (prediction error, where lower values indicate better accuracy) and the right axis plots the demographic parity (DP) gaps for each of the four protected attributes (race, gender, age, and geography, where lower values indicate more equitable outcomes). As the adversarial weight λ increases, all four DP gaps steadily decrease – indicating improved parity across all attributes – while the RMSE correspondingly rises, indicating a degradation in overall predictive accuracy. Notably, these fairness gains begin to plateau after roughly $\lambda \approx 1$, beyond which further increases in λ yield only marginal additional reductions in disparity but incur a steep penalty in RMSE.

6.4 Intersectional debiasing (P6)

A single adversary on race × gender (10 classes) at $\lambda = 2.0$ cuts the intersectional DP to ~ 0.24 (an $\sim 80\%$ reduction) but increases RMSE by $\sim 10\%$. This confirms the trade-off steepens at higher λ.

6.5 RQ1: shape of the frontier

For $\lambda \leq 0.5$, DP_{race} and DP_{geo} drop roughly linearly (~ 0.1–0.2 per step) while RMSE rises $< 1\%$. Beyond $\lambda \approx 1$, fairness gains plateau and accuracy degrades faster—typical Pareto behaviour.

6.6 RQ2: feasible fairness point

Under a single shared λ, no operating point satisfies $\text{DP}_{\text{race}} \leq 0.10$, $\text{DP}_{\text{geo}} \leq 0.50$, and $\Delta\text{RMSE} \leq 5\%$ simultaneously. The nearest is P5 at $\lambda = 1.0$ (all DPs $\downarrow \sim 50\%$) but with $\Delta\text{RMSE} \approx +8.8\%$. *Implication:* a practical path is to combine (i) modest in-processing pressure ($\lambda \approx 0.5$–1.0) with (ii) a light post-processing calibration step (e.g., rebate extremes) targeted at the residual geo bias.

6.7 Discussion and threats to validity

Why gender/age barely moved. Signal for these attributes appears embedded in features/structure not easily removed by a single GRL head; message-passing may re-introduce leakage.
Shared vs. per-attribute λ. A single scalar restricts the Pareto set. A small λ_{geo} increase with a lower λ_{race} could reach a better point.
Synthetic realism. Bias parameters are known and stationary; real networks may have heavy-tailed spatial/temporal dynamics. Phase 7 observed lower RMSE and smaller geo DP than synthetic, indicating qualitative differences and suggesting that λ should be re-tuned for real data.
Small-cell pruning. Intersectional DP excludes groups with < 20 samples, potentially hiding tail effects.
Tuning on subgraphs. Random search on a 10k induced subgraph slightly over-fit; we therefore retained the simpler configuration for full runs.

6.8 Phase 7: Real-World Portfolio (P7)

Using the reduced **15k-node** real-world insurance graph with numeric, categorical, and `City_Code` proximity edges (Section 3), we trained the baseline 3-layer GCN ($\lambda = 0$) with the same protocol (70/15/15 split). On the fixed test set (2,250 samples):

AUC	0.4867	**Accuracy**	0.6169
$\textbf{DP}_{\text{City}}$	0.750	$\textbf{DP}_{\text{Accommodation\_Type}}$	0.066
$\textbf{DP}_{\text{Is\_Spouse}}$	0.025	$\textbf{DP}_{\text{Reco\_Insurance\_Type}}$	0.057

Interpretation. Geography (`City_Code`) shows the largest disparity, with smaller but non-negligible gaps for accommodation type, spouse status, and recommended insurance type. Overall predictive performance is modest (AUC ≈ 0.49), and the higher geo-based gap

compared to synthetic runs suggests that any fairness–accuracy tuning (λ adjustments) would need recalibration for the real portfolio.

6.9 Summary

Multi-attribute adversarial training offers substantial, simultaneous DP improvements, but a shared λ alone cannot hit all fairness targets without overshooting the 5% accuracy budget. The frontier suggests a hybrid strategy (per-attribute λ or mild post-processing) for compliance-ready deployment. Phase 7 confirms this pattern on real data: geography is the primary disparity ($\text{DP}_{\text{City}} \approx 0.75$), with smaller but notable gaps for accommodation type, spouse status, and recommended insurance type. Overall predictive performance remains modest (AUC ≈ 0.49), reinforcing the need for modest in-processing (λ) plus light post-processing to meet fairness targets.

7 Conclusion and Future Work

7.1 Restatement of research question and objectives

This thesis asked whether *multi-attribute adversarial debiasing* inside a Graph Neural Network (GNN) can bring an auto-insurance pricing model within regulatory fairness bounds while keeping the rise in predictive error within 5%. Concretely, we (i) built a relational, bias-controlled synthetic benchmark; (ii) designed a 2-layer GCN with Gradient-Reversal adversaries; (iii) swept λ across single, multi-attribute, and intersectional heads; and (iv) quantified the full fairness–accuracy frontier.

7.2 Key findings

- **Baseline.** The untuned 2×GCN achieved strong accuracy (RMSE $\approx$ 2.62) but large demographic-parity (DP) gaps: race $\approx$ 1.33, geo $\approx$ 1.65; gender $\approx$ 0.05; age $\approx$ 0.11.
- **Single-attribute debiasing (P1–P4).** Race and geo heads reduced their respective DP gaps with $\leq$1% RMSE cost; gender/age moved little.
- **Multi-attribute head (P5).** At $\lambda = 1.0$, all four DP metrics were roughly halved, but RMSE rose by $\sim$8.8%, exceeding the 5% business tolerance.
- **Intersectional head (P6).** A race×gender adversary drove intersectional DP to $\sim$0.24 at $\lambda = 2.0$ but increased RMSE by $\sim$10%.
- **Phase 7 (real portfolio, P7).** On a 15k-node classification graph (City_Code proximity + numeric features), the baseline 3-layer GCN ($\lambda = 0$) achieved AUC ≈ 0.49, Accuracy ≈ 0.62, and $\text{DP}_{\text{City}} \approx 0.75$, with other demographic gaps modest. Geography remains the dominant disparity; recalibration of λ will be required for real-data deployment.

Summary. No single λ satisfied all fairness and accuracy constraints simultaneously; however, the Pareto frontier is now quantified and reproducible.

7.3 Success against objectives

Objective	Status	Evidence
O1 Literature Scan	✓	Reviewed 25+ peer-reviewed works; documented protected attributes, model types, and metrics.
O2 Model Design	✓	2-layer GCN + GRL heads implemented and validated.
O3 Bias-on-Demand Data	✓	`synthetic_nodes.csv`, `synthetic_edges.csv` with bias knobs; bias audit performed.
O4 λ-Sweep Experiments	✓	Tables/plots for P0–P6.
O5 Frontier Analysis (RQ1)	✓	Slopes and trade-off curves.
O6 Feasibility Test (RQ2)	△ Partial	Best $\lambda = 1.0$ halves DPs but breaches 5% RMSE ceiling by $\sim$3.8 pp.
O7 Real-Data Prototype	✓	Resource-constrained 15k-node run (numeric-only, City_Code proximity); qualitative transfer; metrics in Section 6.8.

7.4 Limitations and what they imply

- **Metric scope.** We optimised primarily for *Demographic Parity*. DP is simple and regulator-friendly, but it does not condition on the outcome label, can conflict with calibration, and is sensitive to base-rate differences across groups. As a result, a model that looks fair by DP may still violate *Equalised Odds* (EO) or produce unequal loss-ratios across groups.

- **Single shared λ.** Using one scalar for all adversaries limited control; per-attribute leakage pathways differed (race/geo responsive; gender/age stubborn).

- **Synthetic realism & intersectional sparsity.** The benchmark abstracts roadway topology/temporal dynamics; very small intersectional cells were pruned to avoid unstable estimates.

- **Phase-7 transfer.** Validation executed as a lightweight 15k-node prototype (numeric-only features; `City_Code` proximity edges). Results transfer qualitatively (geography dominates) with **AUC** ≈ 0.49, **Accuracy** ≈ 0.62, and a sizeable geographic gap ($\mathbf{DP}_{\text{City}} \approx 0.75$); this implies λ and thresholds should be recalibrated and richer features restored at full scale.

7.5 Implications for practice and regulatory risk

Even where no single operating point satisfies all targets, the framework is directly usable for *model governance*. A carrier can (i) quantify fairness–accuracy trade-offs before filing, (ii) set *guardrails* (e.g., $\text{DP}_{\text{race}} \leq 0.10$, $\text{DP}_{\text{geo}} \leq 0.50$, $\Delta\text{RMSE} \leq 5\%$), (iii) log metrics for audit, and (iv) deploy alerting when drift pushes metrics outside tolerances. This reduces market-conduct and litigation exposure while building customer and regulator trust.

7.6 Future work

1) Broaden fairness tests beyond DP. Evaluate EO (TPR/FPR parity) and *loss-ratio disparity* alongside DP, reporting all three jointly. EO checks error-rate equity and will reveal residual harms DP can miss; loss-ratio disparity ties directly to actuarial adequacy and business impact. Combine these with calibration tests to guard against fairness–calibration conflicts.

2) Automated selection of λ under constraints. Replace manual sweeps with *constrained, multi-objective hyper-parameter search*: Bayesian optimisation or Hyperband that *minimises* ΔRMSE subject to $\mathrm{DP}_{\mathrm{race}} \leq 0.10$ and $\mathrm{DP}_{\mathrm{geo}} \leq 0.50$. Extend to a vector $\boldsymbol{\lambda} = \{\lambda_{\mathrm{race}}, \lambda_{\mathrm{geo}}, \lambda_{\mathrm{gender}}, \lambda_{\mathrm{age}}\}$ so each adversary gets its own schedule; add a curriculum that increases hard-to-debias heads later in training.

3) Post-processing and calibration. Layer light post-processing on top of the $\lambda = 1.0$ model: reject-option style adjustments or EO-calibration to trim residual geo/race gaps. Because P5 already halves all DPs, small post-hoc moves may meet targets within the remaining RMSE margin.

4) Causal/spatial modelling. Introduce counterfactual/causal structure for geography (e.g., location embeddings that block red-lining proxies) and add roadway topology/seasonality. This targets the stubborn geo gap without paying a large RMSE tax.

5) Real-portfolio scaling & monitoring. Scale Phase-7 from the 15k prototype to the full $\sim$60k portfolio using mini-batch GNN training (e.g., PyG `NeighborLoader`) or sampling; restore informative categoricals; re-tune $\boldsymbol{\lambda}$ to match real-data drift; then operationalise a fairness dashboard (DP, EO, loss-ratio), shadow-pricing, and alerts, with remediation playbooks.

7.7 Contribution to knowledge

We extend adversarial in-processing to a *graph-based* pricing setting with multiple and intersectional protected attributes, and we publish a bias-parameterised relational benchmark. The quantified frontier and diagnostic insights should help both researchers and practitioners navigate accuracy–fairness trade-offs in regulated insurance. We also demonstrate qualitative transfer on real data via a 15k-node portfolio prototype, informing deployment calibration.

References

Baumann, P., Renz, J. and Schmid, S. (2023). Bias-on-demand: A controlled benchmark for fairness experiments, *arXiv preprint arXiv:2302.07890* .

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S., Houde, S. et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, *IBM Journal of Research and Development* **63**(4/5): 4:1–4:15.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization, *Journal of Machine Learning Research* **13**: 281–305.

Berk, R. and Sorenson, S. (2021). Challenges in measuring fairness in insurance pricing, *Risk Analysis* **41**(12): 2225–2237.

Bose, A. and Hamilton, W. L. (2019). Edentity: Fair node embeddings via edge deletion, *NeurIPS Workshop on Graph Representation Learning*.

Chen, J., Li, Y., Wu, L. and Hong, J. (2021). Fairgnn: Fairness-aware node classification on graphs, *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 293–302.

Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S. and Hsieh, C.-J. (2019). Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 257–266.
URL: *https://dl.acm.org/doi/10.1145/3292500.3330925*

Creager, E., Madras, D., Jacobsen, J. and Zemel, R. (2020). Secure fair representation learning, *ICLR Workshop on Responsible AI*.

Dai, L., Wang, J., He, B. and Liu, H. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attributes, *ACM SIGKDD*, pp. 153–163.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. and Pontil, M. (2018). Empirical risk minimization under fairness constraints, *Advances in Neural Information Processing Systems* **31**: 2796–2806.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. (2015). Certifying and removing disparate impact, *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268.

Gallego, J., Puelz, R. and Fierro, R. (2012). Premium disparities in minority zip codes: Evidence from florida auto insurance, *Journal of Insurance Regulation* **31**(3): 45–67.

Grari, B., Valera, I., Nadif, M. and Gasso, G. (2021). Fair auto-encoders for insurance pricing, *AAAI Conference on Artificial Intelligence*, pp. 4369–4377.

Hamilton, W. L., Ying, R. and Leskovec, J. (2017). Inductive representation learning on large graphs, *Advances in Neural Information Processing Systems* **30**: 1024–1034.

Hardt, M., Price, E. and Srebro, N. (2016). Equality of opportunity in supervised learning, *Advances in Neural Information Processing Systems* **29**: 3315–3323.

Hu, L., Ma, X. and Zhu, X. (2020). Fairgbm: Gradient boosting for fairness, *ACM Conference on Fairness, Accountability, and Transparency*, pp. 385–395.

International Association of Insurance Supervisors (2023). Application paper on the use of supervisory material on artificial intelligence and machine learning, *IAIS Publications* .

Karimi, A., Vu, T., Meinshausen, N. and Schölkopf, B. (2021). Fairlooker: A benchmark for causal graph fairness, *NeurIPS Fair AI Workshop*.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks, *arXiv preprint arXiv:1609.02907* .

Korol, J., Ribeiro, M. and Passerini, A. (2022). Fairness via edge dropping in graph neural networks, *ICLR Workshop on Responsible AI*.

Lee, K. and Chau, D. H. (2021). Mitigating bias in graph neural networks with balanced neighbor sampling, *Proceedings of The Web Conference (WWW)*, pp. 540–551.

Lindholm, M., Watson, J. and Zhou, S. (2022). Auditing algorithmic bias in auto-insurance underwriting, *Proceedings of the Casualty Actuarial Society Spring Meeting*, pp. 101–120.

Madras, D., Creager, E., Pitassi, T. and Zemel, R. (2018). Learning adversarially fair and transferable representations, *International Conference on Machine Learning*, pp. 3384–3393.

McFadden, T. and Lee, H. (2020). Actuarial audits of racial bias in california automobile pricing, *North American Actuarial Journal* **24**(4): 567–587.

National Association of Insurance Commissioners (2024). Model bulletin on artificial intelligence and machine learning use, *NAIC Publications* .

OpenML (2025). Insurance (OpenML dataset ID 45064), `https://www.openml.org/d/45064`. Accessed 2025-08-10; publicly available at OpenML.

Powell, B. and Gupta, N. (2019). Pricing fairness in u.s. personal auto: An empirical study, *CAS Annual Meeting*, pp. 1–25.

Smith, T. and Rossi, L. (2021). Fairpricers: Post-hoc reweighting for insurance premium equity, *ICML Workshop on Responsible Machine Learning*, pp. 1–6.

Wald, H., Bengio, Y. and Lacoste, A. (2023). Counterfactual fairness in graph neural networks, *International Conference on Learning Representations* .

Wang, Y. and Kifer, D. (2021). Adversarial multi-attribute fairness in credit scoring, *IEEE International Conference on Big Data*, pp. 1234–1243.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Ctgan: Synthesizing tabular data using gans, *Advances in Neural Information Processing Systems*, Vol. 32, pp. 7345–7356.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. (2013). Learning fair representations, *International Conference on Machine Learning*, pp. 325–333.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R. and Prasanna, V. K. (2020). Graphsaint: Graph sampling based inductive learning method, *International Conference on Learning Representations (ICLR)*.
URL: *https://openreview.net/forum?id=BJe8pkHFwS*

Zhang, B. H., Lemoine, B. and Mitchell, M. (2020). Mitigating unwanted biases with adversarial learning, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* pp. 335–340.

Zhang, X., Smith, J. and Patel, A. (2018). Graph neural networks for auto-insurance claim-frequency prediction, *Insurance: Mathematics and Economics* **78**: 412–425.