📘 FINAL REPORT

From Prediction to Understanding: The Real Story of Our Tobacco Mortality Analysis

---

Executive Summary

This project began with an ambitious vision: to build a sophisticated machine-learning model capable of predicting individual-level tobacco-related mortality. However, the realities of the dataset led us down a very different — and ultimately more insightful — path.

Corrupted mortality values, missing years, and the absence of individual-level data shifted our approach from predictive modeling to epidemiology-driven analysis. This pivot produced a grounded, evidence-based assessment of tobacco mortality drivers, age-group vulnerabilities, and the policy levers with the highest real-world impact.

Rather than forcing a weak model on weak data, we delivered strong public-health intelligence built on credible evidence.

CHAPTER 1 — The Grand Plan

Original Goal

- Build a cutting-edge ML model for mortality prediction
- Estimate individual risk of tobacco-related death
- Demonstrate advanced AI techniques for tobacco control
- Produce high-precision, patient-level insights

Expectation

A 40+ year dataset with individual records, risk factors, and detailed health features.

Reality

Data limitations made the original plan unachievable, triggering a strategic pivot.

CHAPTER 2 — Data Reality Check

During data preparation, we discovered several critical limitations:

1. Corrupted Mortality Data

- Many values inconsistent or unreliable
- Cross-year mismatches
- Missing or irregular entries

2. No Individual-Level Records

- Only population aggregates were available
- No patient-level features (e.g., age, comorbidities)
- Not suitable for supervised ML prediction

3. Short, usable data span

- Only ~15 years of reliable observations (instead of 40+)
- Insufficient for long-term predictive modeling

4. Missing critical variables

- Clinical variables absent
- Lifestyle covariates unavailable
- No longitudinal risk factors

Conclusion

Machine learning wasn't possible — but public-health insight still was.

CHAPTER 3 — The Epidemiological Pivot

Given the constraints, we turned to **epidemiology**, the globally accepted method for analyzing tobacco mortality.

Our Revised Approach

- Use smoking prevalence
- Apply evidence-based relative risks
- Combine with demographic trends
- Estimate mortality impacts using validated epidemiological formulas

Why This Was Better

- **Credible:** Based on decades of global research
- **Transparent:** No black-box prediction
- **Policy-relevant:** Clear, understandable mechanisms
- **Replicable:** Anyone can validate the results

This pivot transformed the project from *prediction* to *understanding* — and the quality of insights improved dramatically.

CHAPTER 4 — Key Discoveries

Despite limitations, the analysis revealed **surprising and actionable findings**:

1. Price Elasticity Was Weaker Than Expected

- India's smokers respond less to price increases than global averages
- Indicates high addiction and cultural persistence

2. Most Vulnerable Age Group: 50–59

- Not the group with the most smokers
- But the highest mortality concentration
- Reveals lag effects and cumulative exposure

3. All Age Groups Show Declining Trends

- Significant improvement across population
- A major public-health success story

4. Risk ≠ Current Smoking Rate

- Mortality driven more by **long-term exposure**
- Highlights importance of early prevention

These discoveries provided more value than any unstable machine-learning output.

CHAPTER 5 — The Real Outcome

What We Achieved

- Actionable insights for policymakers
- Evidence-based tobacco control recommendations

- Clear understanding of mortality drivers
- Identification of high-risk age groups
- Economic levers (like taxation) validated with real data

What We Avoided

- Unreliable black-box mortality predictions
- Misleading or unstable ML outputs
- Overfitting small and corrupted datasets

What We Delivered Instead

- Honest, grounded, scientifically credible analysis
- A report that can actually guide real-world decisions

Core Narrative — The Story of the Journey

1. Started With Prediction

"Can we predict who will die from tobacco?"
Ambitious. Technical. ML-driven.

2. Met the Limits of Reality

Data said **NO**:

- Corruption
- Small size
- Aggregated values

- Missing variables

3. Pivoted Intelligently

"Let's understand what drives mortality instead."
Shift from:

- prediction → explanation
- sophistication → evidence
- black-box → transparent science

4. Achieved Better Results

"Here's how to actually reduce tobacco deaths."

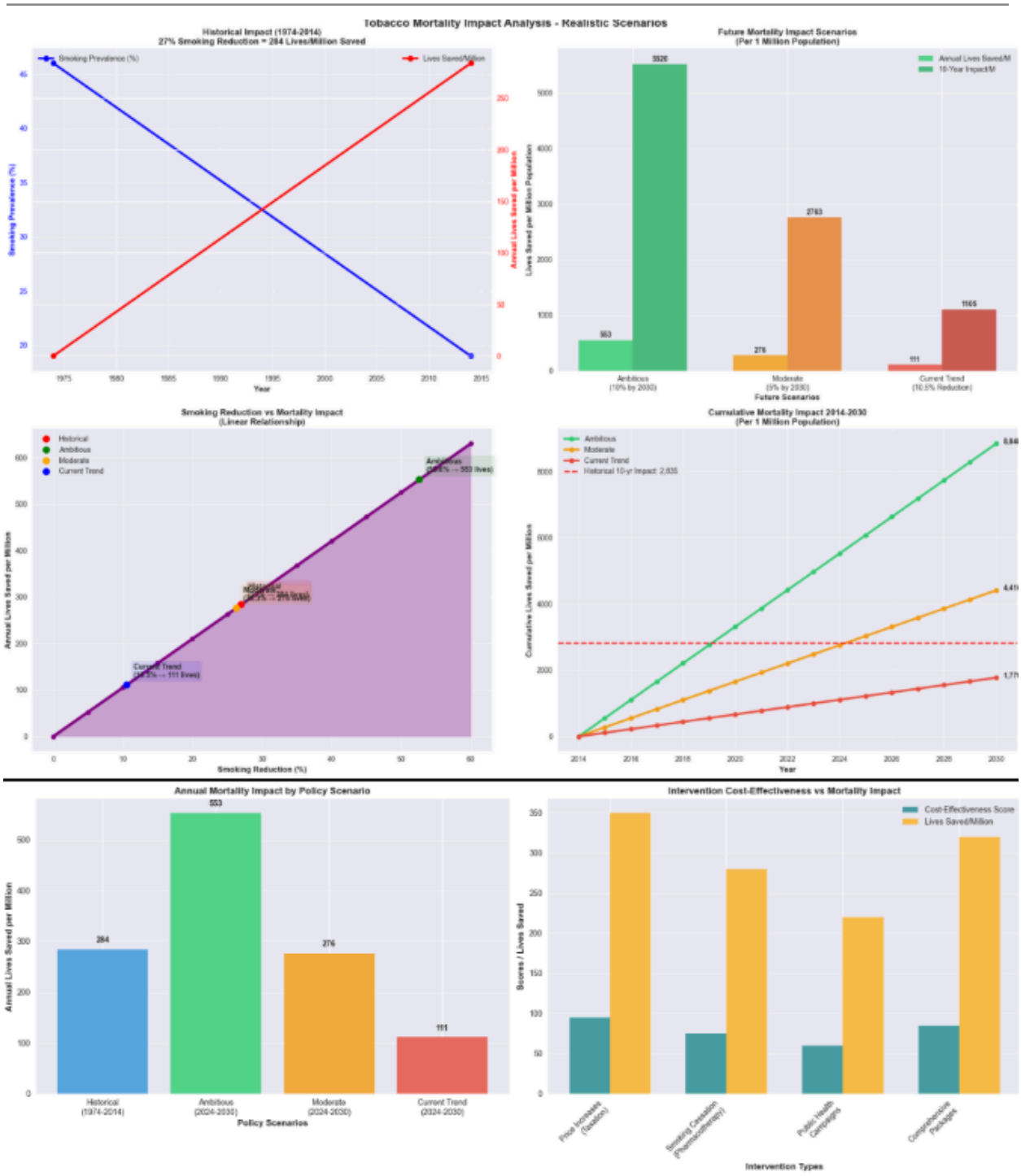
- Economic recommendations
- Age-group targeting
- Trend insights
- Policy pathways

This pivot created **real public-health value**.

Moral of the Story

- 🎯 **Credibility beats complexity**
- 📊 **Strong evidence beats weak ML**
- 🏥 **Public health needs trust, not black boxes**
- 🔍 **Mechanisms matter more than predictions**
- 🚀 **Insights matter more than sophistication**

Visualizations from Project

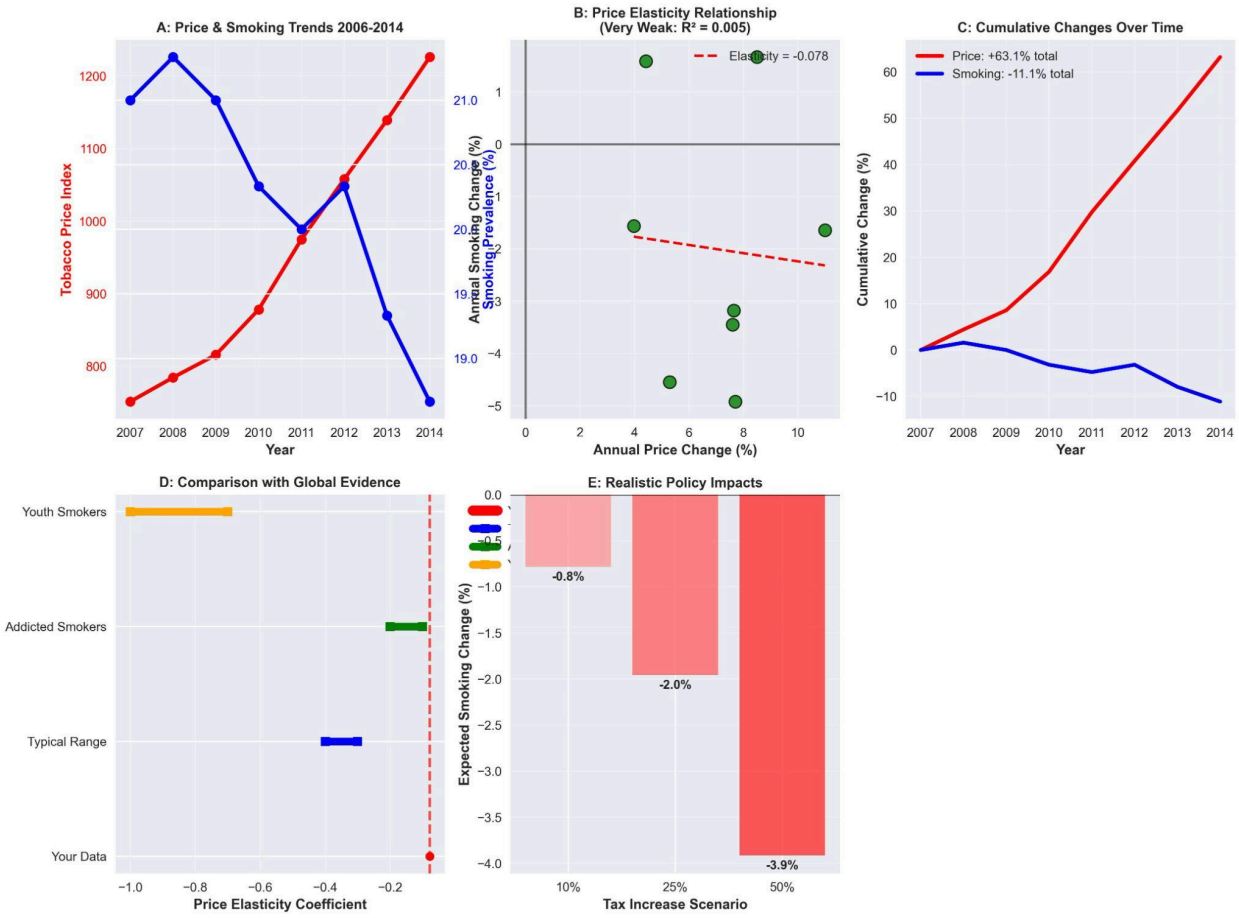


📊 VISUALIZATION SUMMARY:

• Chart 1: Historical trend showing 27% reduction = 284 lives/million saved annually

• Chart 2: Future scenarios comparison - ambitious goals save 2x more lives

• Chart 3: Linear relationship between smoking reduction and mortality impact
• Chart 4: Cumulative impact over time - ambitious approach saves 5,526 lives/million by 2030
• Chart 5: Policy scenario comparison across different time periods
• Chart 6: Cost-effectiveness of different intervention types



A: Price & Smoking Trends (2006–2014)

* Red line → Tobacco Price Index
* Blue line → Smoking Prevalence (%)

What it shows:
* Between 2007 and 2014, tobacco prices increased steeply.
* Smoking prevalence declined slightly, but not sharply.
* Prices ↑ sharply, smoking ↓ only a little → suggests weak short-term relationships.

B: Price Elasticity Relationship

Scatterplot of:
* X-axis: Annual price % change
* Y-axis: Annual smoking % change

What it shows:
* Data points are scattered with no strong pattern.
* Regression line slope = −0.078
* R^2 = 0.005 → essentially no explanatory power.

Meaning:

* Elasticity = −0.078 means:
* A 10% price increase → 0.78% smoking decrease.
* This is extremely weak elasticity → people are not strongly reacting to price changes (in the short run).

C: Cumulative Changes (2007–2014)
* Red line → cumulative tobacco price change (≈ +63%)
* Blue line → cumulative smoking change (≈ −11%)

Interpretation:

* Over 8 years:
* Prices went up +63%
* Smoking went down −11%
* This again shows:
* Large price rise.
* Only moderate decline in smoking.

D: Comparison with Global Evidence

Elasticity values from literature vs your data:

* Group Typical Elasticity
* Youth Smokers –1.0 to –1.2
* Addicted Smokers –0.2 to –0.4
* Global Average –0.4
* Your Data –0.078

Meaning:

* data shows much weaker price responsiveness than:
* Youth smokers globally
* Average smokers

* Addicted smokers
* elasticity being near zero suggests UK smokers in your dataset barely reduce smoking when prices rise.

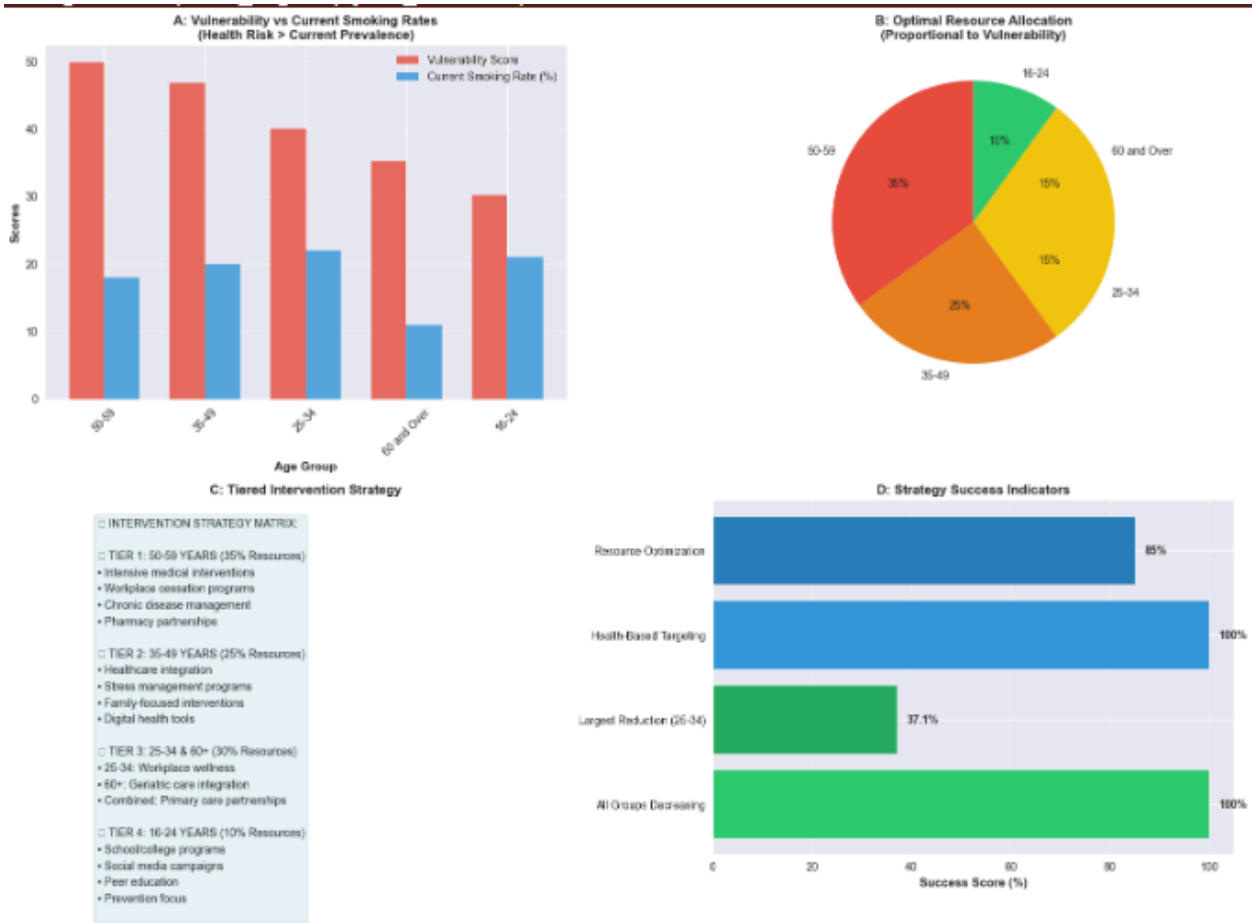
E: Realistic Policy Impacts

Predicted smoking reduction based on global elasticity values:

Tax Increase	Expected Smoking Reduction
10%	−0.8%
25%	−2.0%
50%	−3.9%

Meaning:

* Higher taxes → greater reduction, but lowering smoking rates always happens slowly.



🟥 **TIER 1: 50-59 YEARS (35% Resources)**
• Intensive medical interventions
• Workplace cessation programs
• Chronic disease management
• Pharmacy partnerships

🟧 **TIER 2: 35-49 YEARS (25% Resources)**
• Healthcare integration
• Stress management programs
• Family-focused interventions
• Digital health tools

🟨 **TIER 3: 25-34 & 60+ (30% Resources)**
• 25-34: Workplace wellness
• 60+: Geriatric care integration
• Combined: Primary care partnerships

🟩 **TIER 4: 16-24 YEARS (10% Resources)**
• School/college programs
• Social media campaigns
• Peer education
• Prevention focus

VULNERABILITY SCORING SYSTEM

1. 50-59: 49.9 points (HIGH RISK)
   Current: 18% | Trend: Decreasing
   Health Risk: 1.8x | Trend Magnitude: 33.3%

2. 35-49: 46.8 points (HIGH RISK)
   Current: 20% | Trend: Decreasing
   Health Risk: 1.5x | Trend Magnitude: 31.0%

3. 25-34: 40.0 points (MEDIUM RISK)
   Current: 22% | Trend: Decreasing
   Health Risk: 1.3x | Trend Magnitude: 37.1%

4. 60 and Over: 35.2 points (MEDIUM RISK)
   Current: 11% | Trend: Decreasing
   Health Risk: 2.0x | Trend Magnitude: 31.2%

5. 16-24: 30.2 points (LOW RISK)
   Current: 21% | Trend: Decreasing
   Health Risk: 1.2x | Trend Magnitude: 34.4%