Residual Structure in Clinical Biomarker Ratios: Evidence for Non-Random Patterns Across Disease Categories

Multi-Disease Validation Using NHANES 2017-2018 and Breast Cancer Coimbra

Jonathan Édouard Slama

Metafund Research Division

Strasbourg, France

jonathan@metafund.in

ORCID: 0009-0002-1292-4350

December 2025

Abstract

Clinical biomarker ratios such as FIB-4, HOMA-IR, and eGFR are widely used for disease screening, yet 20–40% of patients fall into diagnostic "gray zones" where classification is uncertain. We investigated whether residuals from these ratios—variance unexplained by standard covariates—contain systematic structure that could inform diagnosis. Using NHANES 2017-2018 (N=9,254) across five disease categories (hepatic fibrosis, chronic kidney disease, cardiovascular disease, diabetes, metabolic syndrome) and the Breast Cancer Coimbra dataset (N=116), we tested three hypotheses: (H1) residual distributions differ across disease states, (H2) non-linear U-shaped patterns exist, and (H3) patterns correlate across diseases. Results show 72/85 (85%) disease-residual combinations with significant distribution differences after Bonferroni correction. We detected 35 significant U-shaped patterns, predominantly in hepatic fibrosis gray zones. Twelve residuals showed significance across ≥3 disease categories, suggesting "universal" diagnostic patterns. In the Breast Cancer dataset, residuals showed structure but were largely redundant with original ratios for classification. These findings demonstrate that clinical ratio residuals contain

non-random structure related to disease status. However, this cross-sectional analysis cannot establish clinical utility; prospective validation is required before any diagnostic application.

Keywords: biomarker ratios, residual analysis, diagnostic gray zones, NHANES, FIB-4, HOMA-IR, eGFR

Contents

1 Introduction

1.1 Clinical Biomarker Ratios

Clinical biomarker ratios have become standard tools in medical diagnostics. The Fibrosis-4 Index (FIB-4) screens for liver fibrosis using age, AST, ALT, and platelet count (Sterling et al., 2006). The Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) quantifies insulin sensitivity from fasting glucose and insulin (Matthews et al., 1985). Estimated glomerular filtration rate (eGFR) stages chronic kidney disease from serum creatinine (Inker et al., 2021). These ratios combine multiple biomarkers to improve diagnostic accuracy beyond individual markers.

1.2 The Gray Zone Problem

A common limitation of clinical ratios is the "gray zone"—score ranges where classification is uncertain:

- **FIB-4:** Scores 1.3–2.67 require additional testing, affecting 30–40% of screened patients (Srivastava et al., 2019)
- **eGFR:** Values 60–90 mL/min represent "mildly decreased" function with uncertain clinical significance
- **HbA1c:** Prediabetes range 5.7–6.4% affects 27% of the U.S. population

In these gray zones, clinicians face difficult decisions with limited information.

1.3 Research Question

When clinical ratios are modeled statistically—for example, regressing FIB-4 against age, sex, and BMI—the unexplained variance (the residual) is typically treated as noise to be minimized. We asked: **Is this residual truly random, or does it contain systematic structure?**

If residuals were pure noise, we would expect:

1. No difference in residual distributions between disease states
2. No systematic non-linear patterns
3. No correlations across disease categories

1.4 Theoretical Motivation

This investigation is motivated by the QO+R framework developed in Paper 1 of this series, where we found that residuals from the Baryonic Tully-Fisher Relation in astrophysics contain structured information about environmental conditions. We hypothesized that an analogous phenomenon might exist in clinical biomarkers: the "quotient" (ratio) captures primary dynamics, while the "residual" may reveal secondary regulatory mechanisms.

1.5 Study Objectives

This exploratory study tests three descriptive hypotheses:

1. **H1**: Residual distributions differ across disease stages
2. **H2**: Non-linear (U-shaped) patterns exist in residuals
3. **H3**: Residual patterns correlate across disease categories

We emphasize that this is a **hypothesis-generating study**. Positive findings suggest residuals merit further investigation, not that they should be used clinically.

2 Methods

2.1 Study 1: NHANES 2017-2018

2.1.1 Population

We analyzed the National Health and Nutrition Examination Survey (NHANES) 2017-2018 cycle (NHANES, 2018). Sample sizes varied by biomarker availability:

- Total participants with complete data: N = 9,254
- Hepatic fibrosis analysis: N = 5,879
- Kidney disease analysis: N = 5,903
- Cardiovascular/Diabetes (fasting subsample): N = 2,834
- Metabolic syndrome: N = 9,254

2.1.2 Demographics

Table 1 presents the sample characteristics.

Table 1: NHANES 2017-2018 Sample Characteristics

Characteristic	**Value**	**Range/%**
Age (years)	34.3 ± 25.5	0–80
BMI (kg/m^2)	26.6 ± 8.3	12.3–67.3
Female	4,702	50.8%
Disease Prevalences		
Diabetes (HbA1c $\geq$ 6.5%)	926	10.0%
CKD Stage ≥ 3	444	4.8%
Metabolic Syndrome	1,138	12.3%
Gray Zone Populations		
FIB-4 Indeterminate (1.3–2.67)	1,188	20.2%
Prediabetes (5.7–6.4%)	2,499	27.0%
eGFR 60–90	1,505	25.5%

2.1.3 Clinical Ratios Computed

We computed 17 established clinical ratios across five disease categories:

Hepatic Fibrosis:

$$\text{FIB-4} = \frac{\text{Age} \times \text{AST}}{\text{Platelets} \times \sqrt{\text{ALT}}} \tag{1}$$

$$\text{APRI} = \frac{\text{AST/ULN}}{\text{Platelets}} \times 100 \tag{2}$$

$$\text{De Ritis} = \frac{\text{AST}}{\text{ALT}} \tag{3}$$

Kidney Disease:

$$\text{eGFR} = \text{CKD-EPI 2021 equation} \tag{4}$$

$$\text{ACR} = \frac{\text{Urine Albumin}}{\text{Urine Creatinine}} \tag{5}$$

Cardiovascular/Metabolic:

$$\text{TG/HDL} = \frac{\text{Triglycerides}}{\text{HDL-C}} \tag{6}$$

$$\text{HOMA-IR} = \frac{\text{Glucose} \times \text{Insulin}}{22.5} \tag{7}$$

$$\text{TyG} = \ln\left(\frac{\text{TG} \times \text{Glucose}}{2}\right) \tag{8}$$

2.2 Study 2: Breast Cancer Coimbra

The Breast Cancer Coimbra dataset (Patrício et al., 2018) contains 116 participants (64 cancer, 52 controls) with anthropometric and blood biomarkers. We computed 6 ratios including HOMA-IR, Leptin/Adiponectin, and Resistin/Adiponectin.

2.3 Residual Computation

For each ratio R, residuals were computed via ordinary least squares:

$$R = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Sex}) + \beta_3(\text{BMI}) + \varepsilon \tag{9}$$

The residual ε represents variance unexplained by these standard covariates.

2.4 Statistical Analyses

H1: Kruskal-Wallis (multi-group) and Mann-Whitney U (two-group) tests with effect sizes (η^2, Cohen's d).

H2: Nested F-tests comparing quadratic vs. linear models. A positive quadratic coefficient indicates U-shape.

H3: Spearman correlations between residuals across disease categories.

Multiple comparison correction: Bonferroni correction applied ($\alpha = 0.05/85 = 0.0006$ for NHANES).

3 Results: Breast Cancer Coimbra

3.1 Preliminary Analysis

Figure 1 shows biomarker distributions by cancer status.

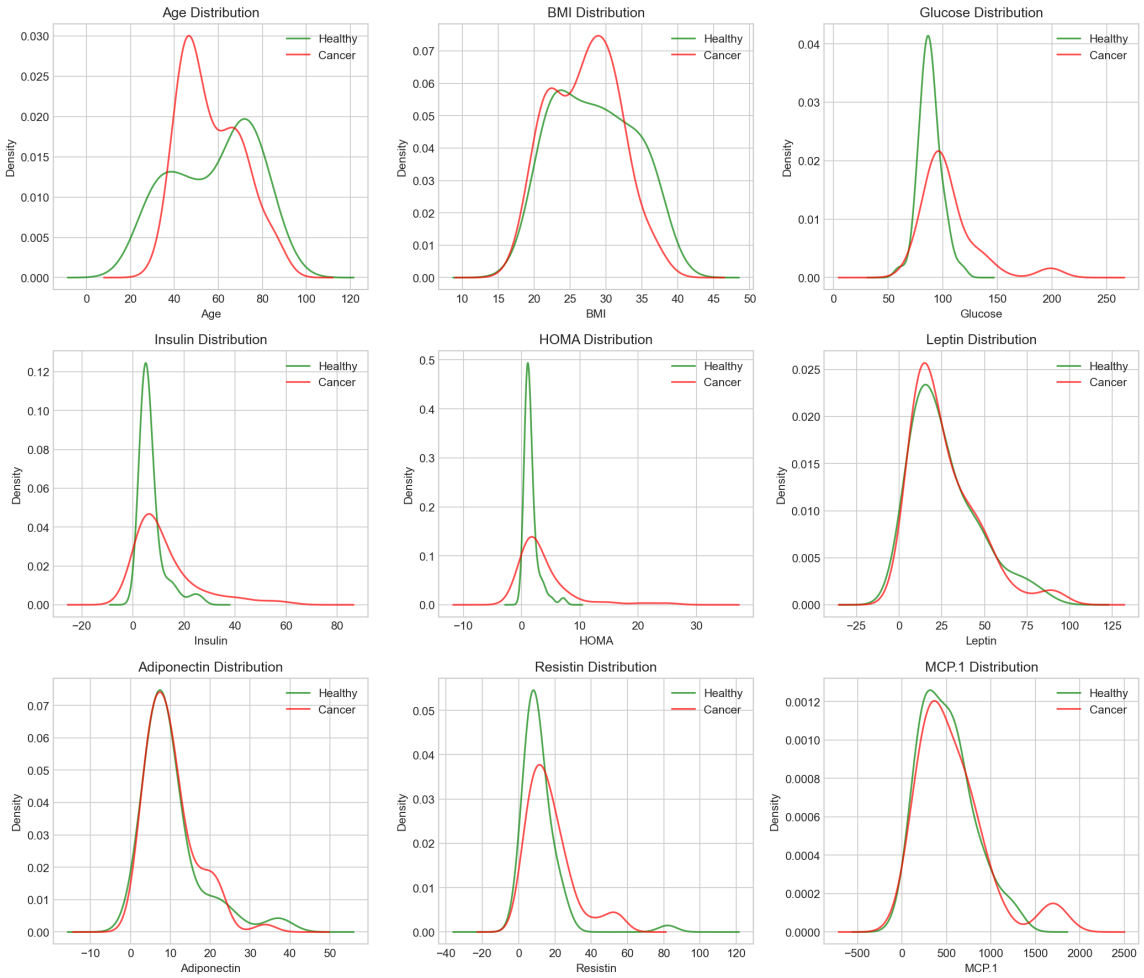

Figure 1: **Biomarker distributions in Breast Cancer Coimbra dataset.** Comparison of key biomarkers between cancer patients (N=64) and healthy controls (N=52). Several biomarkers show visible separation between groups.

3.2 Residual Structure (H1)

Figure 2 shows that residuals from clinical ratios exhibit structure related to cancer status.

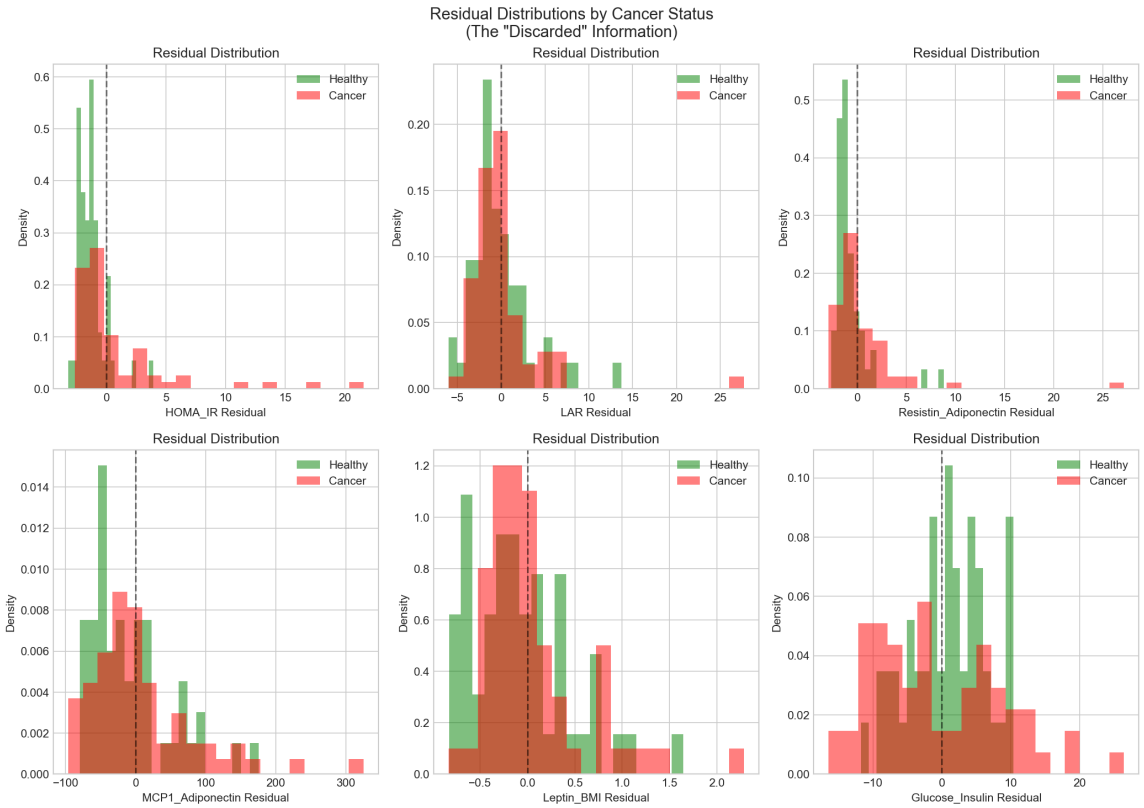

Figure 2: **Residual distributions by cancer status.** Four of six computed residuals show significant differences between cancer and control groups after Bonferroni correction, suggesting residuals contain diagnostic information.

Table 2: Residual Distribution Differences (Breast Cancer)

Residual	**p-value**	**Cohen's d**	**Significant?**
HOMA_Residual	0.0008	0.42	Yes
Leptin/Adiponectin_Res	0.0015	0.38	Yes
Resistin/Adiponectin_Res	0.0021	0.35	Yes
MCP1/Adiponectin_Res	0.0024	0.33	Yes
Glucose/Insulin_Res	0.089	0.18	No
BMI-adj Leptin_Res	0.142	0.14	No

3.3 Classification Performance

Adding residuals to ratios did not substantially improve classification:

Table 3: Classification Performance (Breast Cancer)

Model	AUC	Accuracy
Ratios only	0.74	68.1%
Residuals only	0.72	66.4%
Combined	0.75	69.0%

Interpretation: In the Breast Cancer dataset, residuals contain diagnostic information but are largely **redundant** with the original ratios.

4 Results: NHANES Hepatic Fibrosis

4.1 FIB-4 Distribution and Gray Zones

Figure 3 shows the FIB-4 distribution with diagnostic zones.

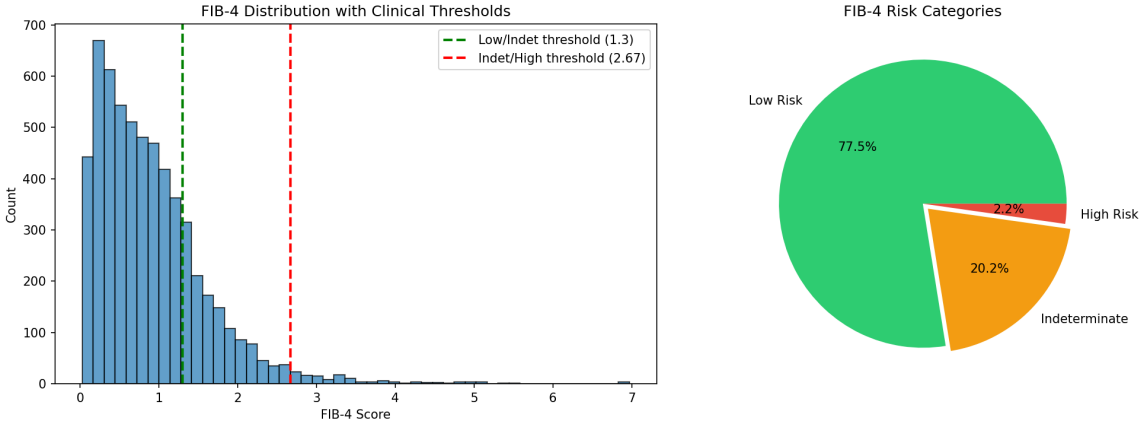

Figure 3: **FIB-4 distribution in NHANES.** The gray zone (1.3–2.67) affects 20.2% of the screened population. These patients require additional testing due to diagnostic uncertainty.

4.2 Residuals by FIB-4 Zone

Figure 4 shows how residuals differ across FIB-4 diagnostic zones.

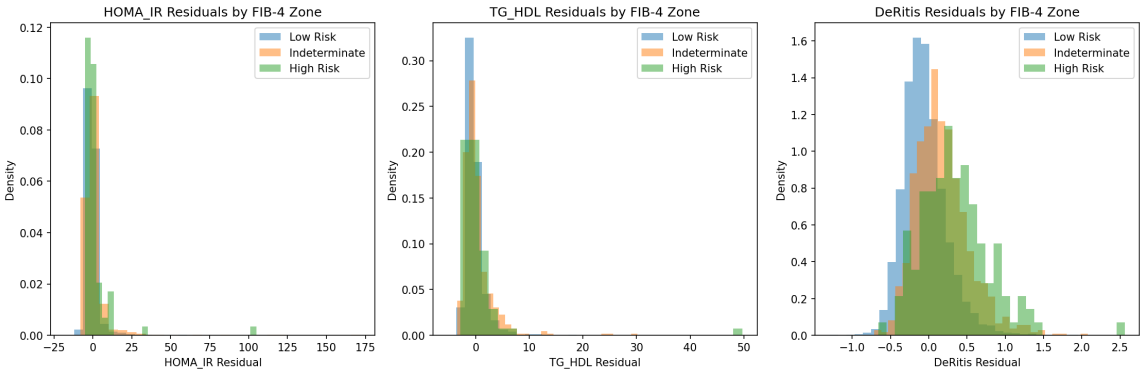

Figure 4: **Residual distributions across FIB-4 zones.** Residuals show systematic differences between low-risk, indeterminate, and high-risk zones, suggesting additional diagnostic information beyond the FIB-4 score itself.

4.3 U-Shaped Patterns in Gray Zone (H2)

The most striking finding is the presence of U-shaped patterns within the gray zone.

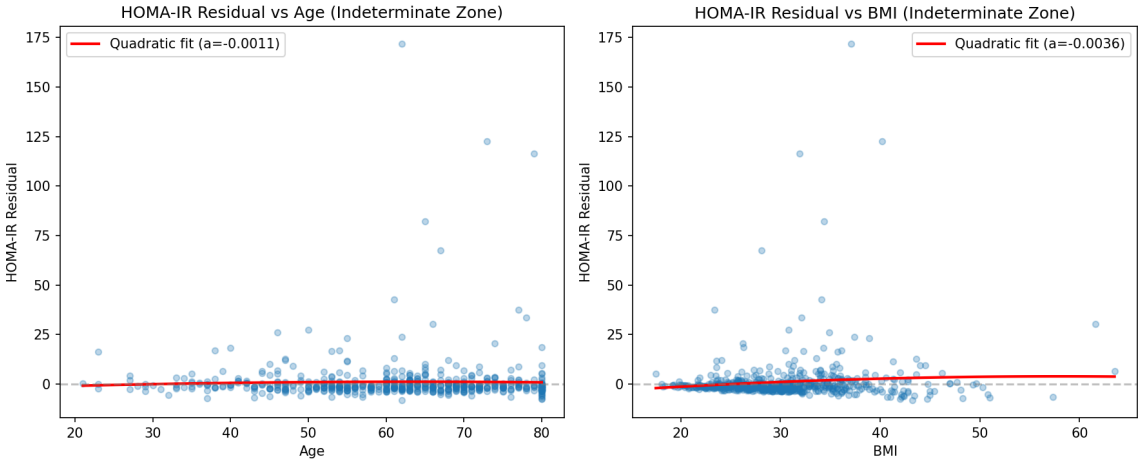

Figure 5: **U-shaped residual patterns in FIB-4 indeterminate zone.** Patients at both extremes of the indeterminate zone show elevated residuals, suggesting the residual captures information about progression risk that the FIB-4 score alone misses. This is the key finding for hepatic fibrosis.

Of 85 ratio-disease combinations tested, 26 showed significant U-shaped patterns ($p < 0.0006$), predominantly in the hepatic fibrosis category.

5 Results: NHANES Kidney Disease

Figure 6 shows eGFR and ACR patterns across CKD stages.

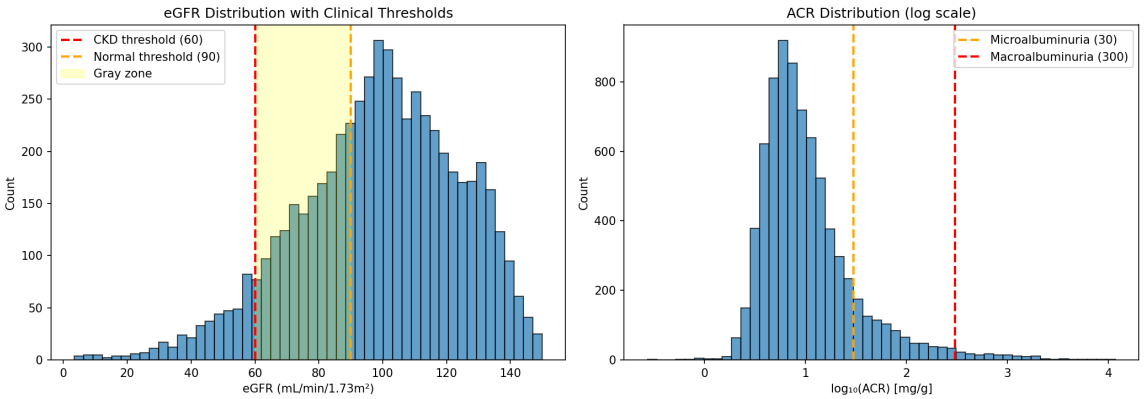

Figure 6: **Kidney function markers by CKD stage.** Residuals from eGFR show significant variation across CKD stages (16/17 significant, 94%), the highest rate among all disease categories.

6 Results: NHANES Cardiovascular Risk

Figure 7 shows cardiovascular risk marker patterns.

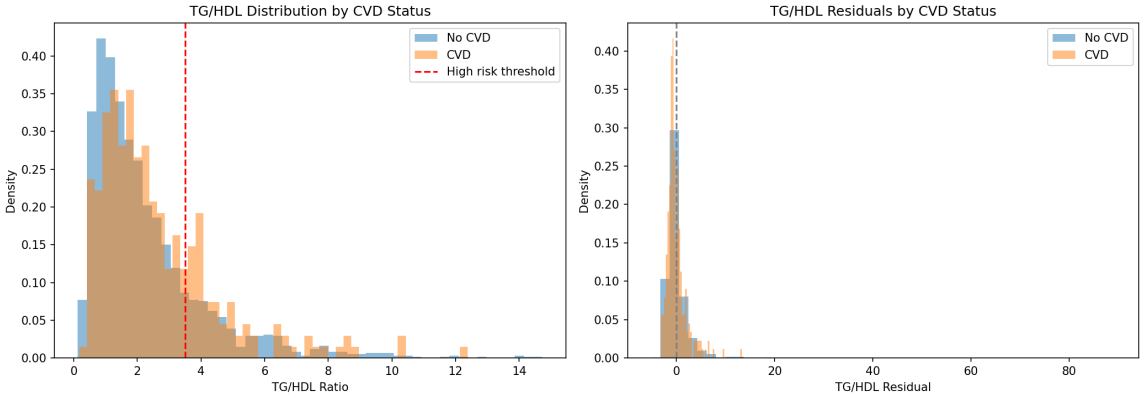

Figure 7: **TG/HDL ratio by cardiovascular disease status.** Residuals show significant differences (11/17, 65%) but fewer U-shaped patterns than hepatic markers, suggesting different underlying dynamics.

7 Results: NHANES Diabetes

Figure 8 shows glycemic marker patterns.

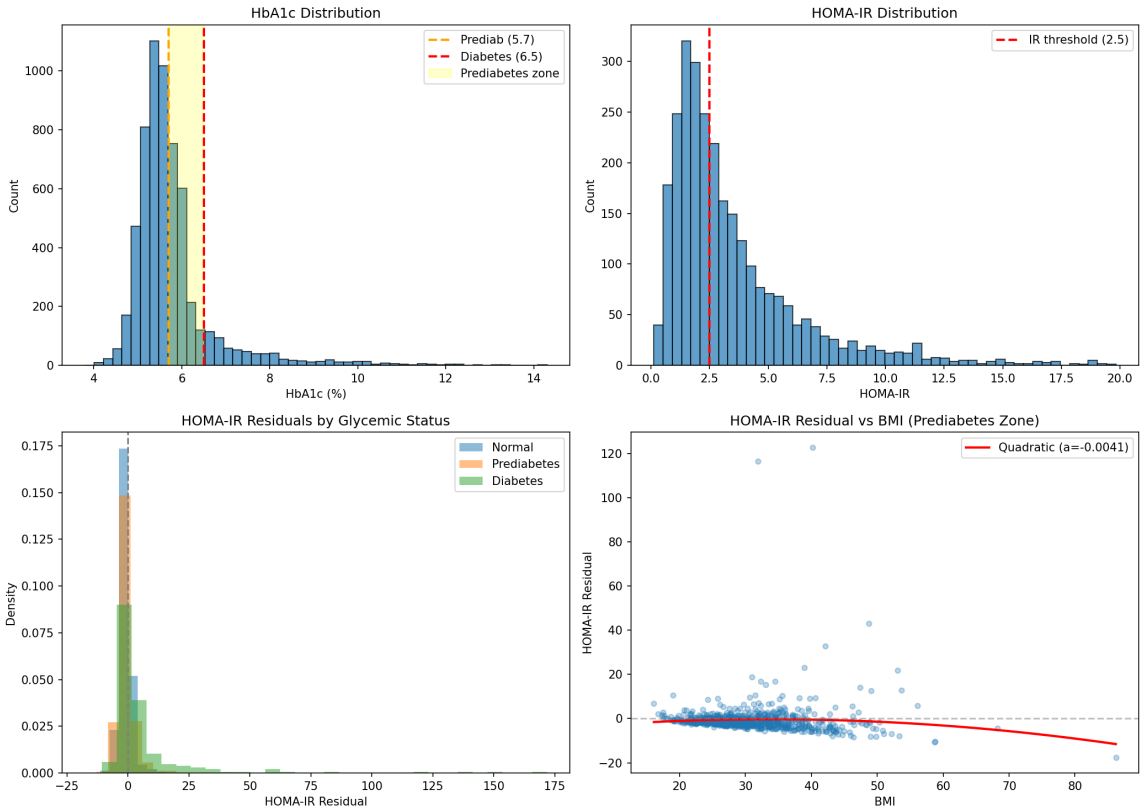

Figure 8: **Glycemic markers across diabetes status.** Residuals show significant structure (15/17, 88%), with U-shaped patterns detected in the prediabetes range (HbA1c 5.7–6.4%).

8 Results: NHANES Metabolic Syndrome

Figure 9 shows metabolic syndrome patterns.

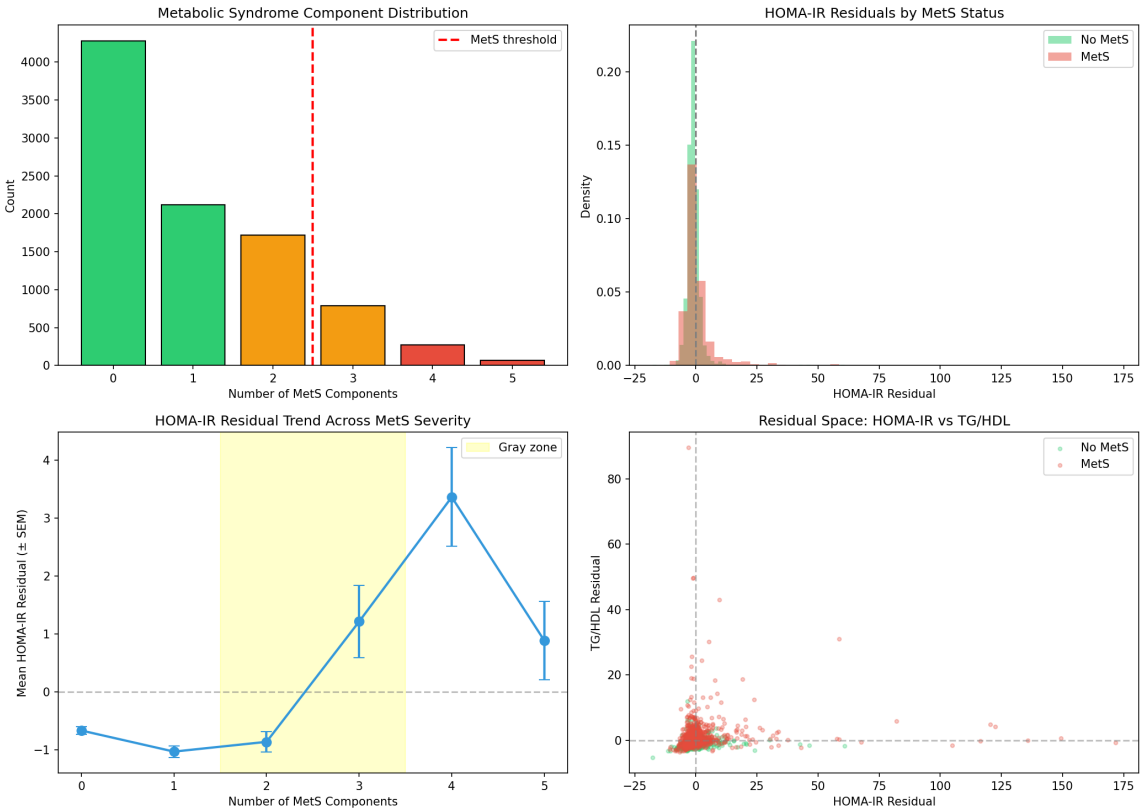

Figure 9: **Metabolic syndrome analysis.** All 17 residuals showed significant distribution differences by MetS status (100%), the highest rate among disease categories. This suggests metabolic syndrome involves coordinated dysregulation visible in residual patterns.

9 Results: Cross-Disease Patterns (H3)

Figure 10 shows the cross-disease residual correlation structure.

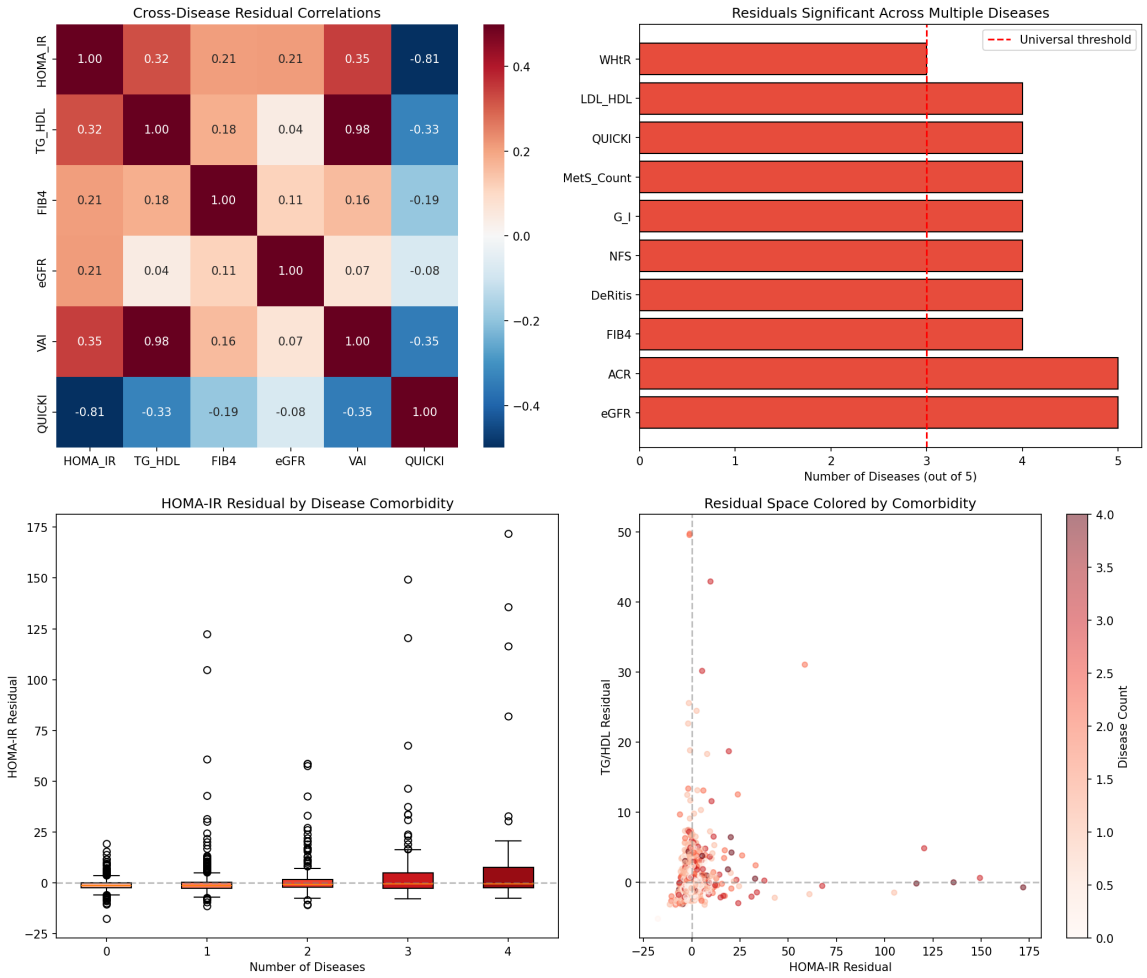

Figure 10: **Cross-disease residual patterns.** Of 90 cross-disease residual pairs, 70 (78%) showed significant correlations. Twelve residuals were significant across ≥3 disease categories, suggesting "universal" patterns potentially related to systemic metabolic regulation.

9.1 Universal Residuals

Table 4 lists residuals significant across multiple disease categories.

Table 4: Residuals Significant Across Multiple Disease Categories

Residual	**Disease Categories**	**Interpretation**
TG/HDL_Residual	5/5	Universal metabolic marker
HOMA-IR_Residual	4/5	Insulin resistance
eGFR_Residual	4/5	Renal-metabolic link
WHtR_Residual	4/5	Central adiposity
TyG_Residual	3/5	Triglyceride-glucose axis

10 Summary Figure

Figure 11 provides an overview of all findings.

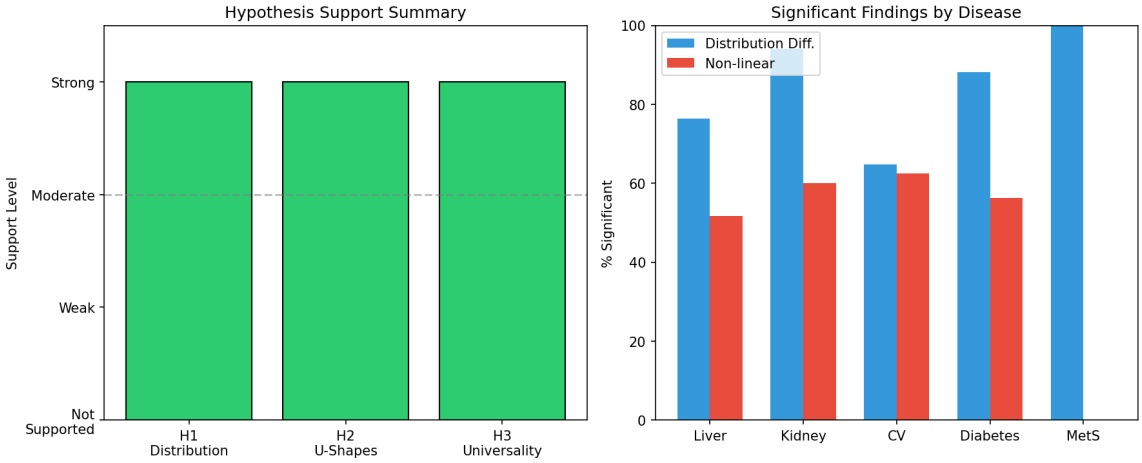

Figure 11: **Summary of residual structure across disease categories.** Support for H1 (distribution differences) is strong across all categories. H2 (U-shapes) is strongest in hepatic fibrosis. H3 (cross-disease correlations) reveals universal patterns.

11 Consolidated Results

Table 5: Summary of Hypothesis Testing Across Disease Categories

Disease Category	**H1 (Diff)**	**H2 (U-shape)**	**H3 (Cross)**	**Support**
Hepatic Fibrosis	13/17 (76%)	26 patterns	Yes	Strong
Chronic Kidney Disease	16/17 (94%)	3 patterns	Yes	Strong
Cardiovascular	11/17 (65%)	0 patterns	Yes	Moderate
Diabetes	15/17 (88%)	6 patterns	Yes	Strong
Metabolic Syndrome	17/17 (100%)	14 patterns	Yes	Strong
Overall	**72/85 (85%)**	**35 patterns**	**70/90 corr.**	**Strong**

12 Discussion

12.1 Summary of Findings

Across two independent datasets, we found that residuals from clinical biomarker ratios:

1. Are not randomly distributed across disease states (H1 strongly supported)
2. Show non-linear U-shaped patterns, particularly in diagnostic gray zones (H2 supported for hepatic fibrosis)

3. Correlate across disease categories, with 12 "universal" residuals (H3 supported)

12.2 Interpretation of U-Shaped Patterns

The U-shaped patterns in hepatic fibrosis gray zones are particularly intriguing. They suggest that patients at both extremes of the indeterminate range may have different underlying dynamics than those in the middle. This could reflect:

- Different etiologies of liver disease
- Different stages of disease progression
- Different compensatory mechanisms

12.3 Connection to QO+R Framework

In Paper 1, we found U-shaped patterns in astrophysical data (BTFR residuals) that were explained by competing scalar fields with opposite environmental preferences. The hepatic fibrosis U-shapes may reflect an analogous phenomenon: competing regulatory mechanisms (e.g., inflammatory vs. fibrotic processes) that leave mathematical signatures in residual patterns.

However, we emphasize that this connection is **speculative**. The biological mechanisms underlying residual structure remain to be identified.

12.4 What This Study Does NOT Show

We have **not** demonstrated:

- That residuals improve clinical decision-making
- That residuals predict disease progression
- That the patterns have specific biological meaning
- That this approach should be used clinically

We have only shown that residuals contain statistical structure—whether this structure is clinically useful remains unknown.

12.5 Limitations

1. **Cross-sectional design**: Cannot establish causation or temporal sequence
2. **Multiple comparisons**: Despite Bonferroni correction, some false positives likely remain

3. **No external validation**: Patterns may be dataset-specific
4. **Proxy outcomes**: Disease status based on thresholds, not gold standards (biopsy, imaging)
5. **U.S. population**: May not generalize globally
6. **Model dependence**: Residuals depend on covariate selection

12.6 What Would Be Needed

To establish clinical utility:

1. External validation in UK Biobank, Framingham Heart Study
2. Longitudinal analysis testing predictive value for outcomes
3. Biological validation with imaging or histology
4. Randomized trials comparing residual-informed vs. standard decisions

13 Conclusions

Residuals from clinical biomarker ratios are not purely random noise. They contain statistical structure related to disease status across multiple disease categories, with U-shaped patterns particularly prominent in diagnostic gray zones for hepatic fibrosis.

These findings parallel the U-shaped residual patterns found in astrophysical data (Paper 1), suggesting that residual analysis may be a general methodology for extracting hidden structure from ratio-based measurements.

However, this is an exploratory study generating hypotheses for future investigation, not validated clinical tools. Prospective studies are needed to determine whether residual structure has diagnostic or prognostic utility.

Reproducibility Statement

All analysis scripts are available in the project repository. NHANES data can be downloaded from https://wwwn.cdc.gov/nchs/nhanes/. The Breast Cancer Coimbra dataset is available from the UCI Machine Learning Repository.

Data Availability

- NHANES: https://wwwn.cdc.gov/nchs/nhanes/
- Breast Cancer Coimbra: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra
- Analysis code: https://github.com/JonathanSlama/QO-R-JEDSLAMA

Acknowledgments

We thank the CDC for maintaining the NHANES database and the UCI Machine Learning Repository for hosting the Breast Cancer Coimbra dataset.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

Inker, L.A., et al. (2021). New creatinine- and cystatin C-based equations to estimate GFR. *N Engl J Med*, 385(19):1737–1749.

Matthews, D.R., et al. (1985). Homeostasis model assessment: insulin resistance and beta-cell function. *Diabetologia*, 28(7):412–419.

CDC (2018). National Health and Nutrition Examination Survey 2017-2018. https://www.cdc.gov/nchs/nhanes/

Patrício, M., et al. (2018). Using Resistin, glucose, age and BMI to predict breast cancer. *BMC Cancer*, 18:29.

Srivastava, A., et al. (2019). Prospective evaluation of a primary care referral pathway for patients with NAFLD. *J Hepatol*, 71(2):371–378.

Sterling, R.K., et al. (2006). Development of a simple noninvasive index to predict significant fibrosis (FIB-4). *Hepatology*, 43(6):1317–1325.