■ THYROID DISEASE UNIFIED DATASET

225,568 Patient Records	**3 Source Datasets**	**37 Variables**
Total Records	Unified & Cleaned	Features

This report presents a comprehensive analysis of a unified thyroid disease dataset assembled from three independent clinical and epidemiological sources. It covers patient demographics, thyroid hormone levels, risk factors, diagnostic outcomes, and a detailed variable dictionary.

February 2025 | Data Science Division

Table of Contents

1. Dataset Overview & Sources

This dataset is the result of a careful unification of three independent thyroid-related databases: **cancer_risk** (an epidemiological dataset focused on thyroid cancer risk factors), **thyroidDF** (a clinical dataset with detailed thyroid function tests), and **hypothyroid** (a specialized dataset for hypothyroidism diagnosis). After cleaning and logical harmonization, the combined dataset contains **225,568 records** and **37 features**.

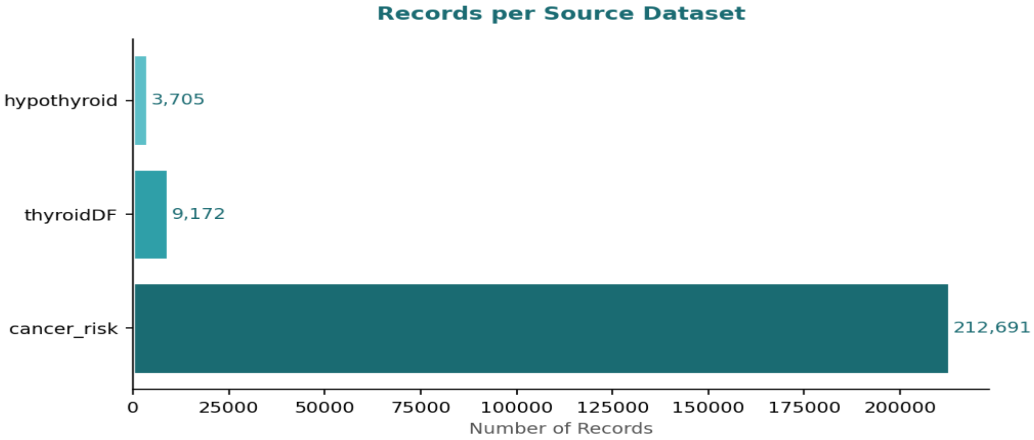

Figure 1 — Number of records contributed by each source dataset.

Source	Records	Primary Focus	Key Variables
cancer_risk	212,691	Cancer risk epidemiology	country, ethnicity, risk factors, nodule_size, diagnosis, class (Low/Medium/High)
thyroidDF	9,172	Thyroid function tests	TSH, T3, TT4, T4, T4U, FTI, medications, query flags
hypothyroid	3,705	Hypothyroidism classification	Age, sex, hormone levels, clinical conditions, class labels

Note: Variables that exist only in specific sources may show high missing rates in the unified dataset — this is expected and reflects the structural differences between the original studies.

2. Patient Demographics

The dataset covers a diverse international patient population spanning all adult age groups. Female patients are slightly over-represented, consistent with the known higher prevalence of thyroid disorders in women.

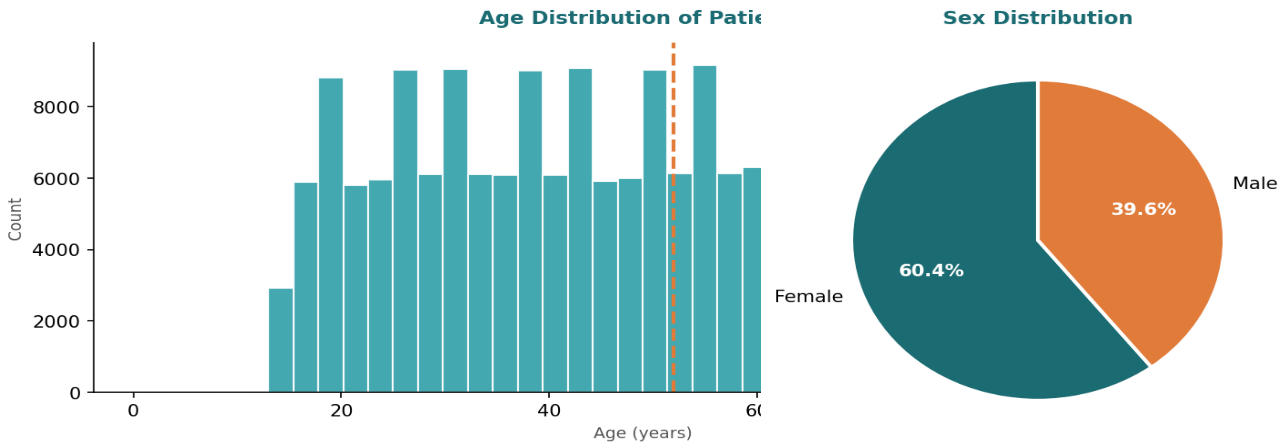

Figure 2 — Age distribution (left) and sex split (right).

Age Summary Statistics

Min	Q1 (25%)	Median	Mean	Q3 (75%)	Max
1	33	52	51.9	70	97

Geographic & Ethnic Distribution

The cancer_risk subset (94% of total records) provides rich geographic diversity, covering patients from across Asia, Africa, Europe, and the Americas. India, China, and Nigeria are the three most represented countries.

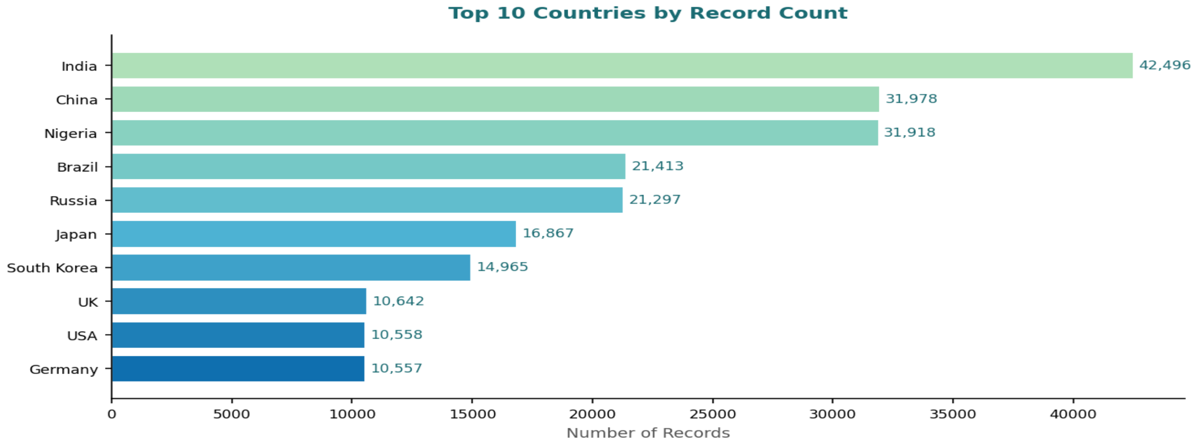

Figure 3 — Top 10 countries by patient record count.

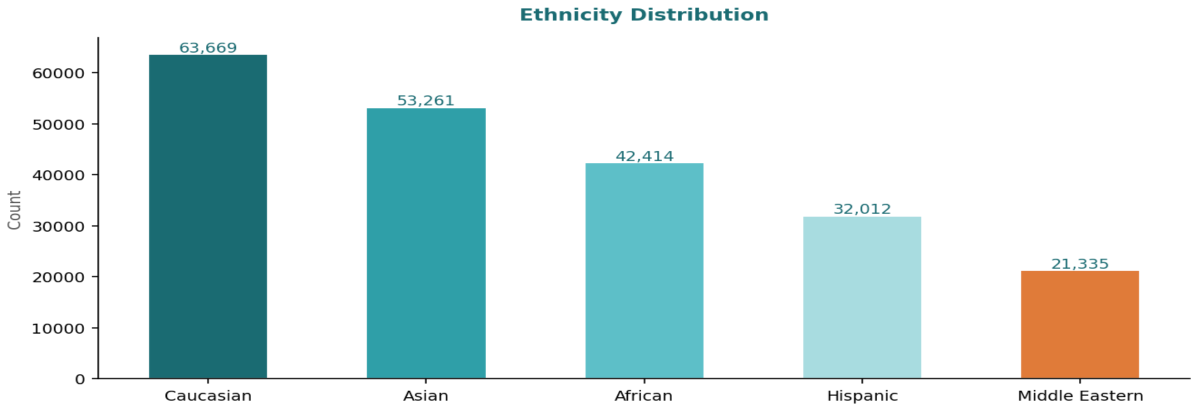

Figure 4 — Ethnicity breakdown across all patients.

3. Thyroid Hormone Measurements

Thyroid hormone assays are the cornerstone of thyroid disease diagnosis. The dataset includes five key laboratory measurements, primarily from the thyroidDF and hypothyroid subsets. These values allow classification of hyperthyroid, hypothyroid, and euthyroid states.

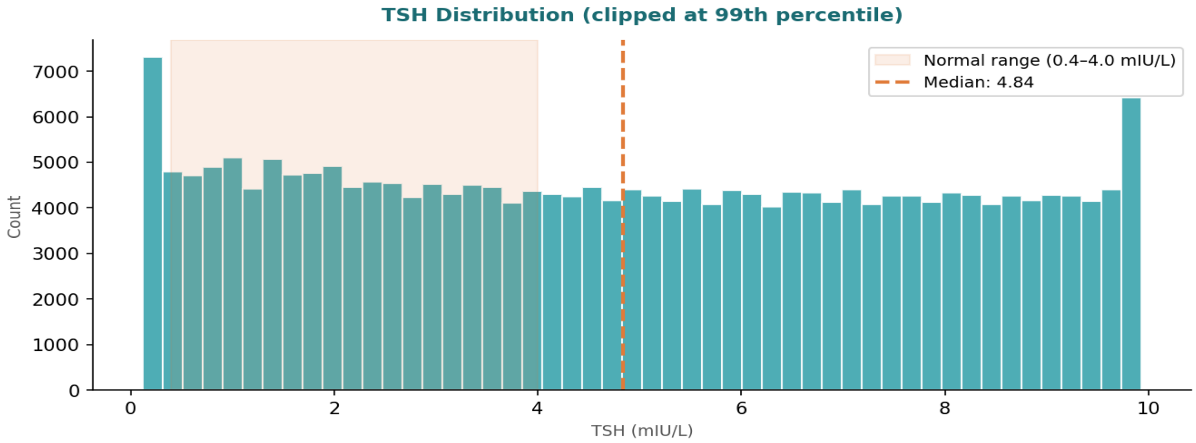

Figure 5 — TSH distribution. The shaded region marks the clinical normal range (0.4–4.0 mIU/L). The median TSH of 4.84 mIU/L suggests many patients fall at or slightly above normal.

Hormone Assay Summary

Hormone	Unit	Normal Range	N (non-null)	Mean	Median	Std
TSH	mIU/L	0.4 – 4.0	224,421	4.90	4.84	2.92
T3	nmol/L	1.2 – 3.1	222,259	2.00	2.00	0.86
TT4	nmol/L	58 – 161	8,730	108.3	104.0	35.0
T4	µg/dL	4.5 – 12.5	216,396	N/A	N/A	N/A
T4U (T4 Uptake)	%	24 – 37	11,745	0.980	0.960	0.193
FTI (Free T4 Index)	index	72 – 163	11,754	111.9	108.0	33.6

4. Risk Factors & Clinical Conditions

Several systemic and environmental risk factors for thyroid disease are captured. These binary (Yes/No) variables apply primarily to the cancer_risk subset. Obesity and family history are the most prevalent comorbidities, each present in roughly 30% of patients.

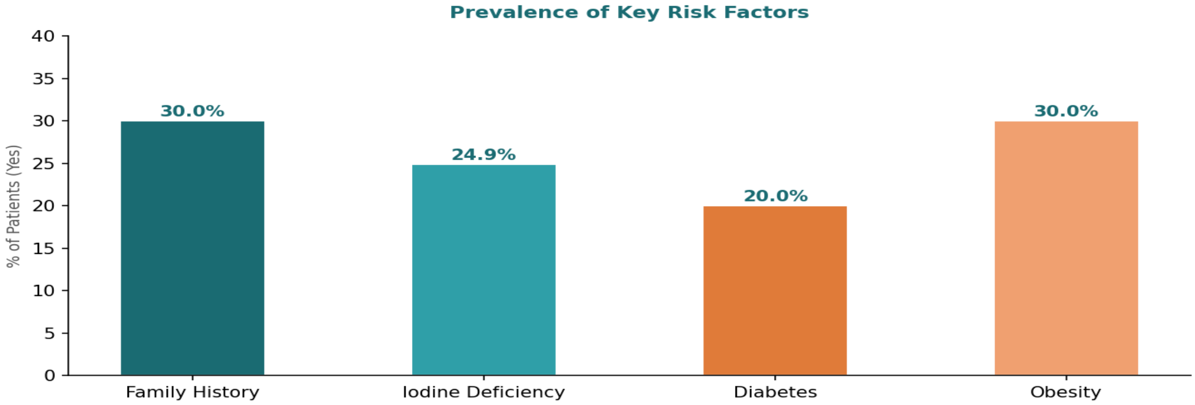

Figure 6 — Prevalence of key risk factors among patients with available data.

Risk Factor	Yes	No	% Yes
Family History	63,825	148,866	30.0%
Iodine Deficiency	53,018	159,673	24.9%
Diabetes	42,593	170,098	20.0%
Obesity	63,886	148,805	30.0%

5. Diagnostic Outcomes

The dataset encodes outcomes at two levels: the **class** variable (primary target), which captures cancer risk tier (Low/Medium/High) or clinical thyroid diagnosis; and **diagnosis**, which provides a direct Benign/Malignant label for the cancer_risk subset.

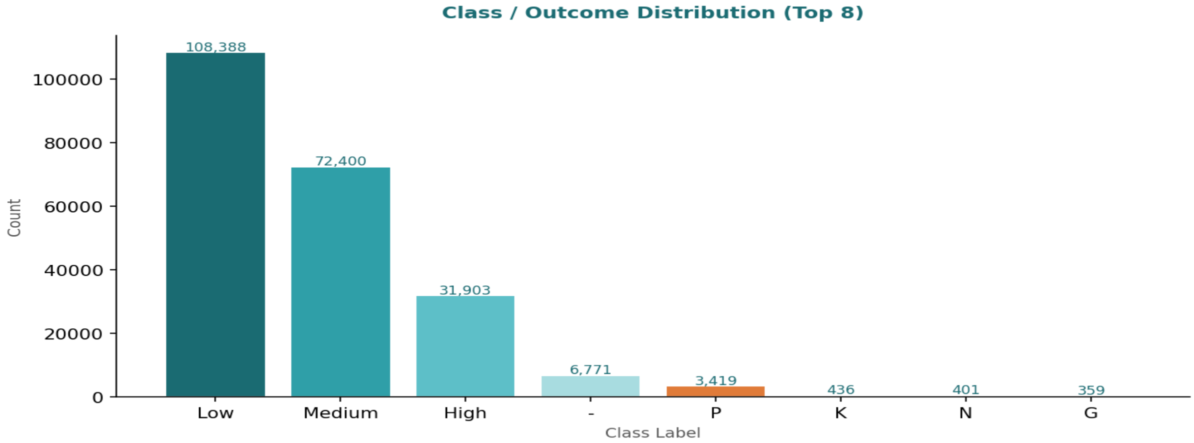

Figure 7 — Distribution of the class variable (top 8 values).

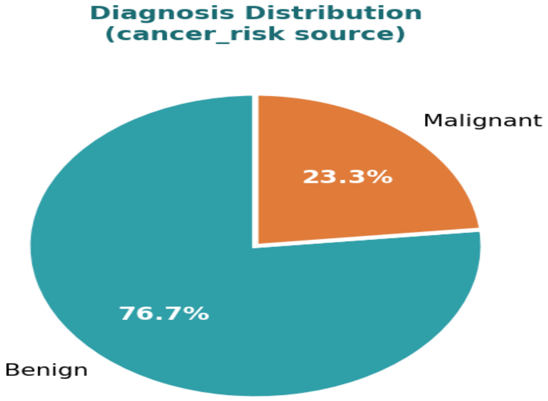

Figure 8 — Benign vs. Malignant split among patients with nodule diagnosis. 23.2% of assessed nodules were malignant.

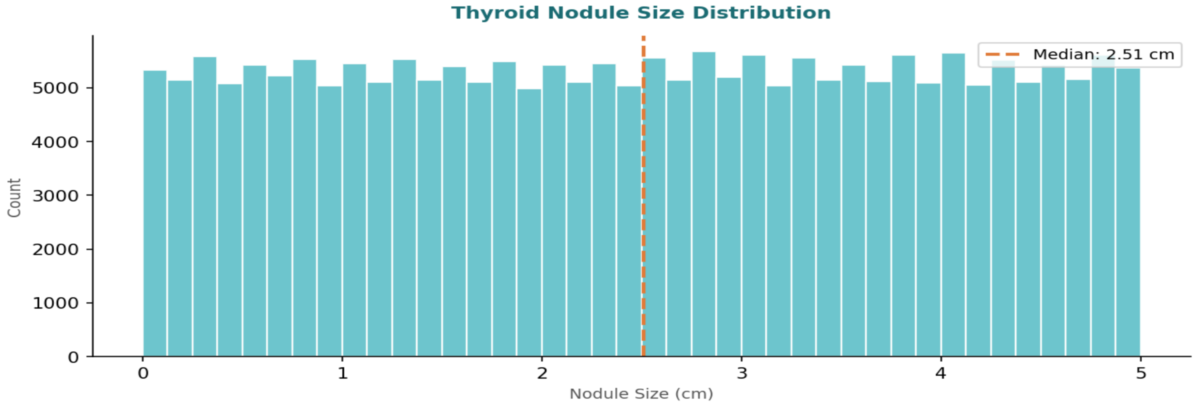

Figure 9 — Distribution of thyroid nodule sizes. The median nodule is 2.51 cm, with sizes ranging from 0 to 5 cm.

6. Data Quality & Missing Values

Missing data is a structural feature of this unified dataset, arising from the different variable scopes of each source. Variables specific to the cancer_risk source (e.g., country, ethnicity, nodule_size) are absent for the ~12,877 records from the other sources. Similarly, variables specific to thyroidDF or hypothyroid (e.g., measured flags, lithium, psych) are missing for the 212,691 cancer_risk records. This missingness is **informative by source** and should be handled with source-aware imputation or subset analysis.

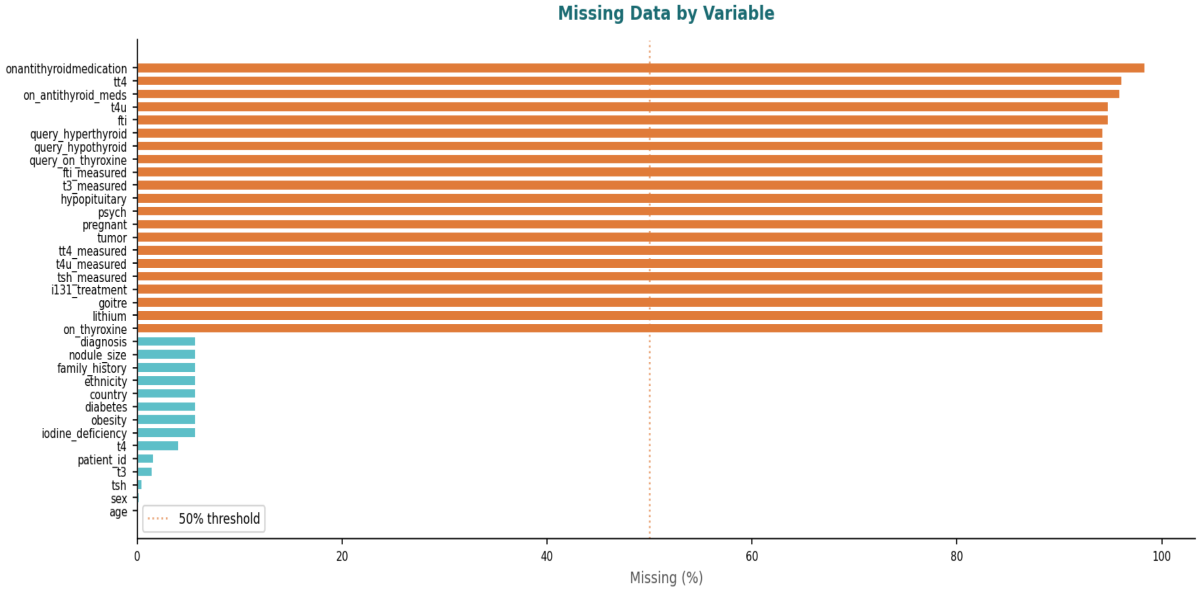

Figure 10 — Missing data percentage per variable. Variables in orange exceed 50% missingness, primarily because they only exist in one source dataset.

7. Variable Dictionary

Below is a complete description of all 37 variables in the unified dataset.

7.1 Identifiers & Metadata

Variable	Type	Description	Source
patient_id	Float / ID	Unique patient identifier. Values differ in format across sources (integer counters in cancer_risk, larger numeric IDs in thyroidDF/hypothyroid).	All
source	Categorical	Identifies the original dataset the record came from. Values: cancer_risk, thyroidDF, hypothyroid. Critical for source-aware analysis.	All

7.2 Demographics

Variable	Type	Description	Source
age	Numeric (years)	Patient age in years. Ranges from 1 to 97; median is 52. Approximately normally distributed, slightly left-skewed.	All
sex	Binary (F/M)	Patient biological sex. Female = 60.4%, Male = 39.6%. Reflects the known higher thyroid disease prevalence in women.	All
country	Categorical (string)	Country of origin or treatment. 10 countries represented. Top: India (42,496), China (31,978), Nigeria (31,918). Available only in cancer_risk.	cancer_risk
ethnicity	Categorical (string)	Self-reported or assigned ethnicity. Five groups: Caucasian, Asian, African, Hispanic, Middle Eastern. Available only in cancer_risk.	cancer_risk

7.3 Risk Factors (Binary: Yes / No)

Variable	Type	Description	Source
family_history	Binary (Yes/No)	Whether the patient has a family history of thyroid cancer or thyroid disease. Present in 30.0% of records with data. Strong hereditary risk factor.	cancer_risk
iodine_deficiency	Binary (Yes/No)	Whether the patient lives in or has been exposed to an iodine-deficient environment. Found in 24.9% of cases. Iodine deficiency is a major global cause of thyroid disorders.	cancer_risk
diabetes	Binary (Yes/No)	Presence of a diabetes diagnosis. Found in 20.0% of patients. Type 2 diabetes is associated with altered thyroid hormone metabolism.	cancer_risk

obesity	Binary (Yes/No)	Whether the patient is classified as obese. Present in 30.0% of patients. Obesity is linked to both hypothyroidism and thyroid cancer risk.	cancer_risk
lithium	Binary (f/t)	Whether the patient is on lithium treatment. Lithium inhibits thyroid hormone secretion and is a known cause of drug-induced hypothyroidism.	thyroidDF / hypothyroid
psych	Binary (f/t)	Whether the patient has a psychiatric condition or is under psychiatric care. Psychological comorbidities are sometimes associated with thyroid dysfunction.	thyroidDF / hypothyroid
pregnant	Binary (f/t)	Whether the patient is currently pregnant. Pregnancy significantly alters thyroid function requirements; hypothyroidism in pregnancy can harm fetal development.	thyroidDF / hypothyroid
goitre	Binary (f/t)	Presence of goitre (enlargement of the thyroid gland). Often caused by iodine deficiency or autoimmune thyroid disease.	thyroidDF / hypothyroid
hypopituitary	Binary (f/t)	Whether the patient has hypopituitarism (underactive pituitary gland). Since the pituitary produces TSH, hypopituitarism leads to secondary hypothyroidism.	thyroidDF / hypothyroid
tumor	Binary (f/t)	Presence of a pituitary or other relevant tumor. Pituitary tumors can disrupt TSH secretion and thyroid function.	thyroidDF / hypothyroid

7.4 Nodule & Cancer-Specific Variables

Variable	Type	Description	Source
nodule_size	Numeric (cm)	Size of the thyroid nodule in centimeters. Ranges from 0 to 5 cm; median 2.51 cm. Larger nodules (>4 cm) carry higher malignancy risk. Available only in cancer_risk.	cancer_risk
diagnosis	Binary (Benign/Malignant)	Pathological diagnosis of the thyroid nodule. 76.8% Benign, 23.2% Malignant. This is the clinical ground truth for cancer classification tasks. Available only in cancer_risk.	cancer_risk

7.5 Medications & Treatments (Binary)

Variable	Type	Description	Source
on_thyroxine	Binary (f/t)	Whether the patient is currently taking thyroxine (levothyroxine, T4 replacement therapy). The most common treatment for hypothyroidism.	thyroidDF / hypothyroid
on_antithyroid_meds	Binary (f/t)	Whether the patient is on antithyroid medications (e.g., methimazole, propylthiouracil). Used to treat hyperthyroidism.	thyroidDF / hypothyroid
onantithyroidmedication	Binary (f/t)	A second encoding of antithyroid medication use (from a different source). May overlap with on_antithyroid_meds; should be reconciled before modeling.	thyroidDF / hypothyroid
i131_treatment	Binary (f/t)	Whether the patient has received radioactive iodine (I-131) treatment. Used to treat hyperthyroidism and thyroid cancer by selectively destroying thyroid tissue.	thyroidDF / hypothyroid

7.6 Thyroid Hormone Measurements

Variable	Type	Description	Source
tsh	Numeric (mIU/L)	Thyroid-Stimulating Hormone. Produced by the pituitary to stimulate the thyroid. High TSH → hypothyroidism; Low TSH → hyperthyroidism. Normal range: 0.4–4.0 mIU/L.	All
t3	Numeric (nmol/L)	Triiodothyronine — the biologically active thyroid hormone. Elevated in hyperthyroidism. Normal range: 1.2–3.1 nmol/L.	All
tt4	Numeric (nmol/L)	Total Thyroxine (T4). Includes both bound and free T4. Normal range: 58–161 nmol/L. Mainly available in thyroidDF/hypothyroid.	thyroidDF / hypothyroid
t4	Numeric (µg/dL)	Free or total T4 level (exact measurement type varies by source). Low in hypothyroidism, high in hyperthyroidism.	thyroidDF

t4u	Numeric (ratio)	T4 Uptake — measures the proportion of free T4. Used in calculating the Free Thyroxine Index (FTI). Normal range: approximately 0.85–1.15.	thyroidDF / hypothyroid
fti	Numeric (index)	Free Thyroxine Index — calculated as TT4 × T4U. A derived measure of free T4 activity. Normal range: approximately 72–163.	thyroidDF / hypothyroid

7.7 Measurement Flags (Binary: t / f)

These boolean flags indicate whether the corresponding hormone was actually measured for that patient. When a measurement flag is False (f), the hormone value is typically missing. These are useful for feature engineering or understanding measurement patterns.

Variable	Type	Description	Source
tsh_measured	Binary (t/f)	Whether TSH was measured for this patient visit.	thyroidDF / hypothyroid
t3_measured	Binary (t/f)	Whether T3 was measured for this patient visit.	thyroidDF / hypothyroid
tt4_measured	Binary (t/f)	Whether TT4 was measured for this patient visit.	thyroidDF / hypothyroid
t4u_measured	Binary (t/f)	Whether T4 uptake was measured for this patient visit.	thyroidDF / hypothyroid
fti_measured	Binary (t/f)	Whether FTI was computed/measured for this patient visit.	thyroidDF / hypothyroid

7.8 Clinical Query Flags (Binary: t / f)

These flags represent clinical suspicions or referral reasons documented at the time of the visit.

Variable	Type	Description	Source
query_hyperthyroid	Binary (t/f)	Whether hyperthyroidism was clinically suspected at the time of referral or test ordering. A diagnostic hypothesis flag.	thyroidDF / hypothyroid
query_hypothyroid	Binary (t/f)	Whether hypothyroidism was clinically suspected. Often used as a triage signal prior to lab confirmation.	thyroidDF / hypothyroid
query_on_thyroxine	Binary (t/f)	Whether the physician queried or suspected the patient was on thyroxine therapy, possibly as an undocumented or unverified treatment.	thyroidDF / hypothyroid

7.9 Target Variable

Variable	Type	Description	Source
class	Categorical (multi-label)	Primary outcome variable. Encodes cancer risk tier (Low, Medium, High) from cancer_risk, and clinical thyroid diagnosis codes from thyroidDF/hypothyroid: "-" = negative/normal, P = primary hypothyroid, K = compensated hypothyroid, G = goitre, I = increased binding protein, F = T3 toxic, R = T3 toxic goitre, A = discordant test result, L = low T4, M = primary hypothyroid (child), S = sick, N = no condition found.	All

8. Key Insights & Recommendations

■ Dataset Scale

With 225,568 records and 37 features, this is a substantial dataset for training machine learning models. The cancer_risk source dominates (94.3%), so models should be evaluated on source-stratified splits.

■ Geographic Diversity

The global scope (10 countries, 5 ethnic groups) is a strength for building generalizable models. However, country and ethnicity are only available for the cancer_risk subset.

■ Hormone Measurements

TSH is the most complete hormone variable (99.5% available). T3 is also well-covered. TT4, T4U, and FTI are only available for ~5% of records (thyroidDF + hypothyroid). Feature engineering should create derived hormone ratios (e.g., T3/TSH).

■■ Structural Missingness

High missingness in variables like on_thyroxine, lithium, psych, etc. is fully explained by dataset source — these variables simply do not exist in cancer_risk. Source-conditional imputation or multi-task modeling is recommended.

■ Target Variable

The class variable encodes two different scales: a 3-level risk tier (cancer_risk) and 15+ clinical thyroid condition codes (thyroidDF/hypothyroid). Analysts should define a unified target schema before training predictive models.

■■ Malignancy Rate

23.2% of assessed nodules were malignant — a significant event rate suitable for binary classification tasks with or without SMOTE balancing.

■ Duplicate Risk Factors

on_antithyroid_meds and onantithyroidmedication appear to encode the same information from different sources. These should be merged or deduplication logic applied.

Data Science Note: This unified dataset provides a strong foundation for multi-task thyroid disease modeling. Recommended next steps include: (1) source-stratified cross-validation, (2) unified class label schema, (3) source-aware imputation pipeline, and (4) feature selection across the three variable groups (demographics, hormones, risk factors).