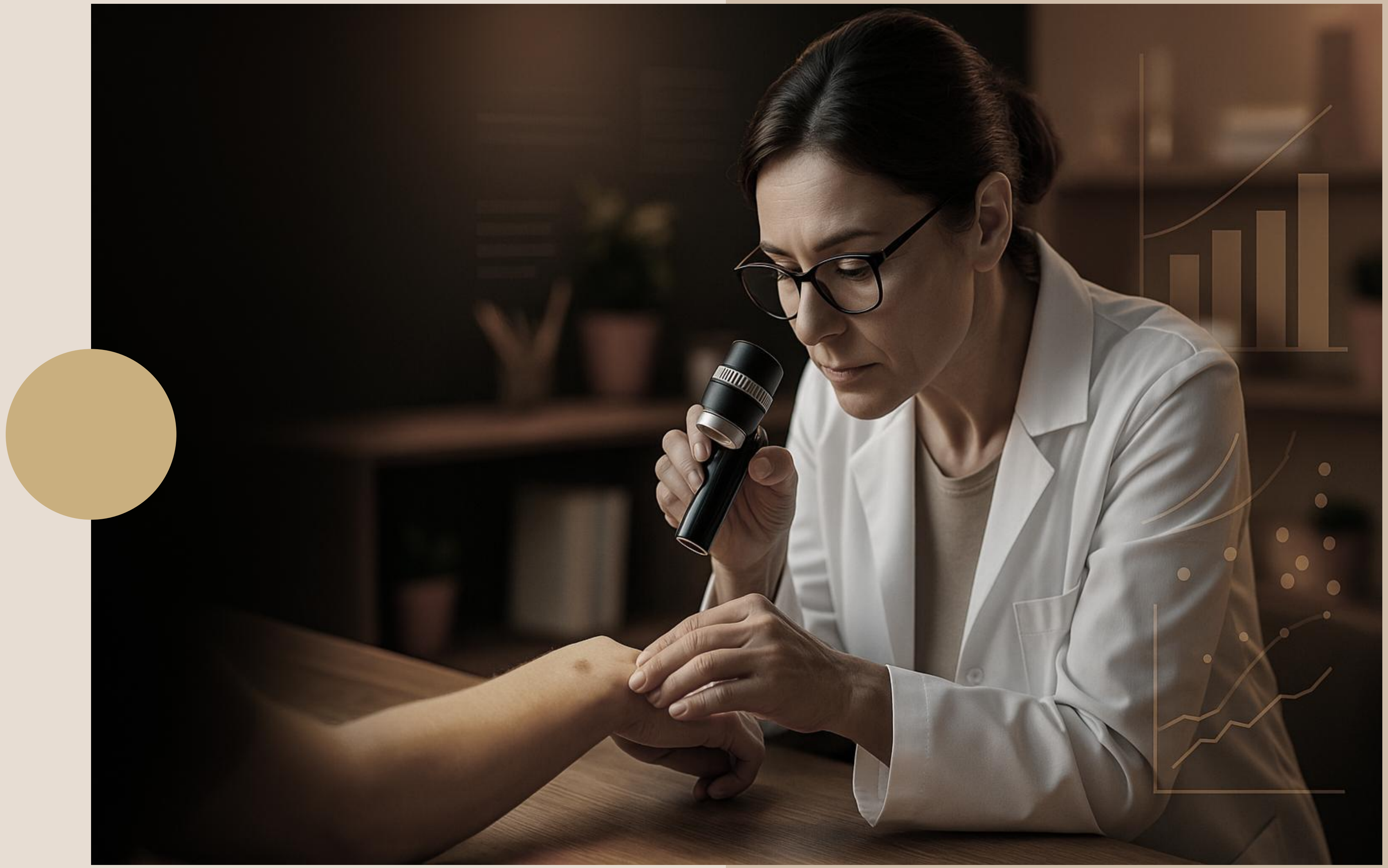
DermAI Diagnostics: SQL Analytics for Early Skin-Cancer Detection



JUSTINA AGBLO

Problem Statement

- Delays in detection stem from misdiagnosis, limited dermatology access, and incomplete understanding of environmental risks.
- With 1,089 skin-lesion instances, we explore links among demographics, environmental exposure, and lesion traits.
- Goal: strengthen early-stage diagnosis and ML-based decision support by structuring the data for SQL analysis and model training.

Data Description

- **patient_id** Unique identifier for each patient
- **smoke** Patient smokes (TRUE/FALSE)
- **drink** Patient drinks alcohol (TRUE/FALSE)
- **background_father** Patient’s paternal ethnicity
- **background_mother** Patient’s maternal ethnicity
- **age** Age of patient
- **pesticide** Exposure to pesticides (TRUE/FALSE)
- **gender** Gender (MALE/FEMALE)
- **skin_cancer_history** Previous skin cancer diagnosis (TRUE/FALSE)
- **cancer_history** Family history of cancer (TRUE/FALSE)
- **has_piped_water** Access to piped water (TRUE/FALSE)
- **has_sewage_system** Access to sewage system (TRUE/FALSE).
- **lesion_id** Unique identifier for each lesion
- **patient_id** Foreign key linking to Patient_Info
- **fitspatrick** Fitzpatrick skin type (1-6)
- **region** Body region of the lesion
- **diameter_1** Diameter of lesion (mm)
- **diameter_2** Second diameter measurement (mm)
- **diagnostic** Type of skin lesion (BCC, MEL, NEV, etc.)
- **itch** Lesion causes itching (TRUE/FALSE)
- **grew** Lesion has grown (TRUE/FALSE)
- **hurt** Lesion causes pain (TRUE/FALSE)
- **changed** Lesion changed in color/size (TRUE/FALSE)
- **Bleed** Lesion bleeds (TRUE/FALSE)
- **elevation** Lesion is raised (TRUE/FALSE)
- **img_id** Associated lesion image filename
- **biopsed** Whether the lesion was biopsy-confirmed (TRUE/FALSE)

Rationale

Bridging Data and Medicine — practical impact for clinicians and patients.

SQL Learning Opportunity — real queries on realistic clinical/lesion data.

AI-Driven Medical Research — prepare ML-ready datasets responsibly.

Early Detection & Prevention — prioritize timely, accurate diagnosis.

Real-World Application — insights usable by healthcare teams.

Core Questions

- Which demographics (age, sex, etc.) correlate with lesion types?
- How do environmental exposures (e.g., UV index, pesticides, smoking, alcohol) relate to cancer risk?
- Which lesion characteristics best separate cancerous vs. benign?
- What patterns support early detection and triage?

Cases & Malignancy by Age Band

Volume skews older:

- 60+ accounts for 50.9% of all cases (554/1088); 45–59 adds 30.2% (329). Together, 45+ = 81.1% of lesions.

Risk climbs with age:

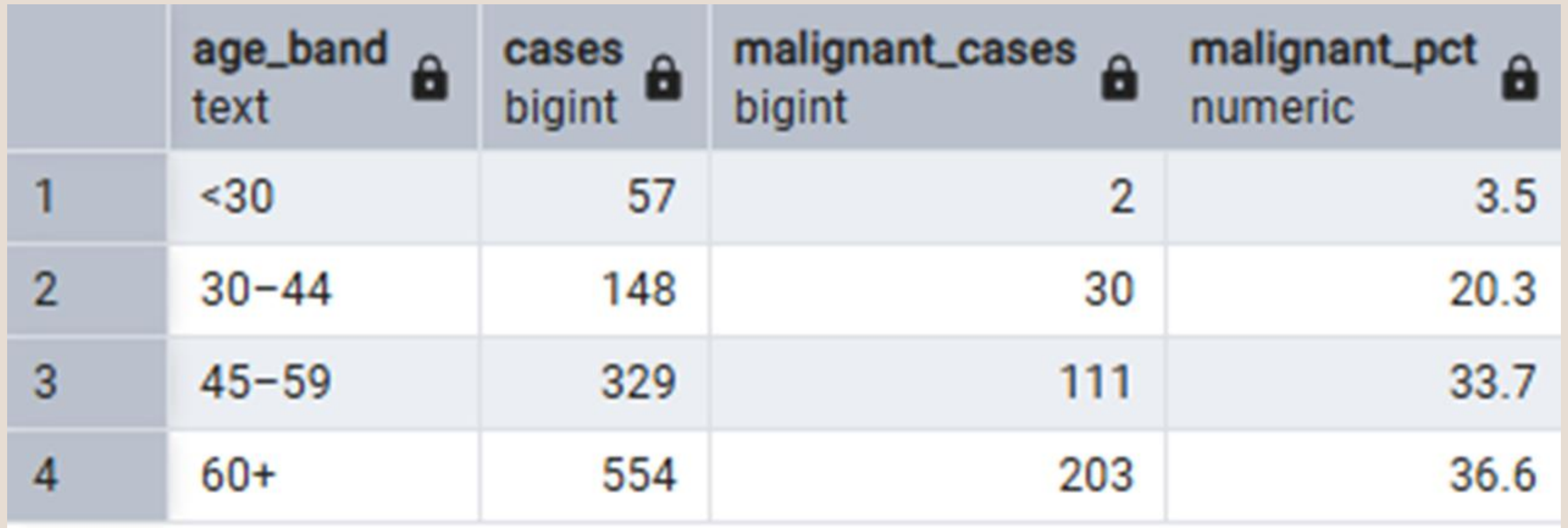
- malignancy rate rises from 3.5% (<30) → 20.3% (30–44) → 33.7% (45–59) → 36.6% (60+) (~10× higher in 60+ vs <30).

Where malignancies actually occur:

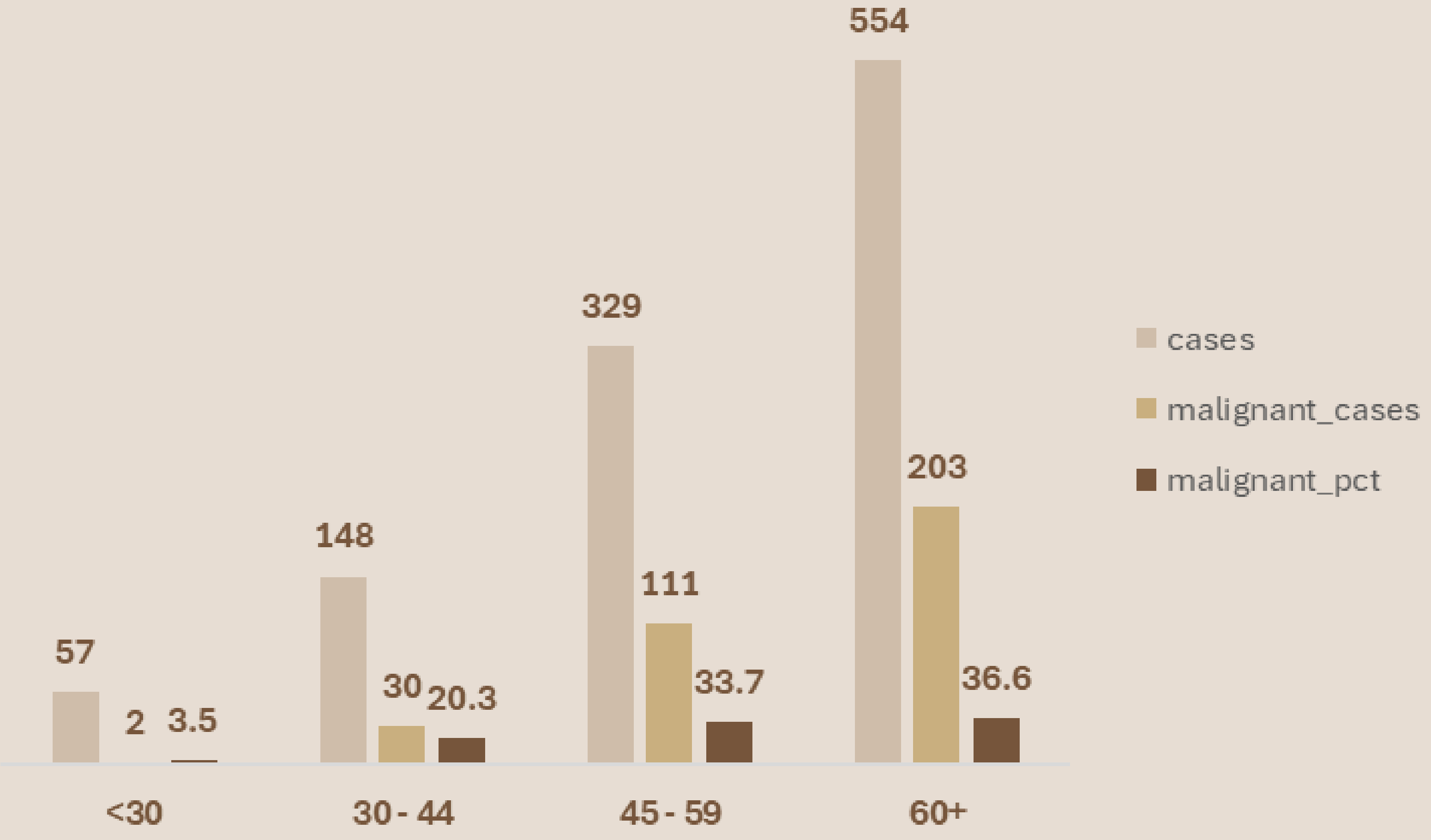
- of 346 malignant cases, 60+ contributes 58.7%, 45–59 = 32.1% — 90.8% are in 45+.

High-volume / high-risk bands:

- 60+ (High/High); 45–59 (High/High).Moderate band: 30–44 (Moderate volume, mid risk).Low-yield band: <30 (Low volume, very low risk).



	age_band text	cases bigint	malignant_cases bigint	malignant_pct numeric
1	<30	57	2	3.5
2	30–44	148	30	20.3
3	45–59	329	111	33.7
4	60+	554	203	36.6



Cases & Malignancy by Sex/Gender

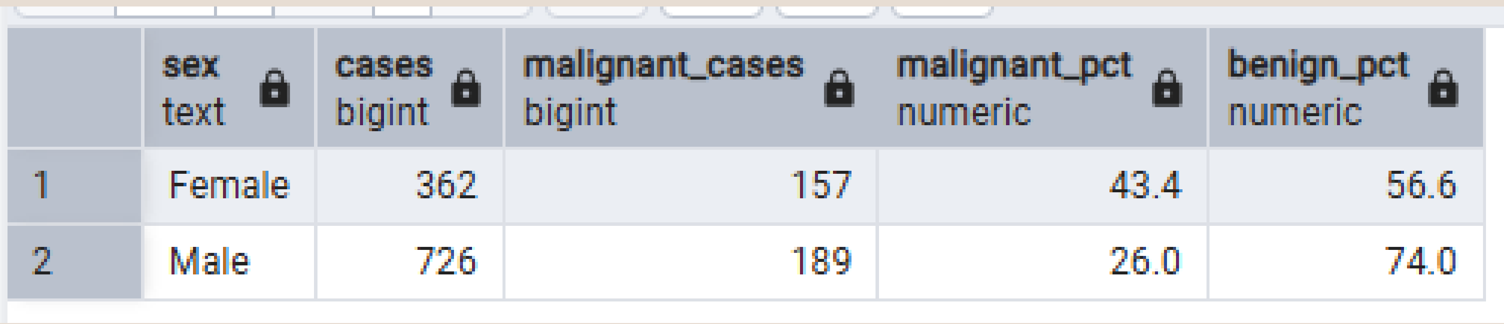
- **Volume:** Males = 726 (66.7%) of cases; Females = 362 (33.3%).

- **Risk:** Malignancy rate is 43.4% in females vs 26.0% in males → females have +17.4 pp higher risk and ~1.67× higher relative risk.

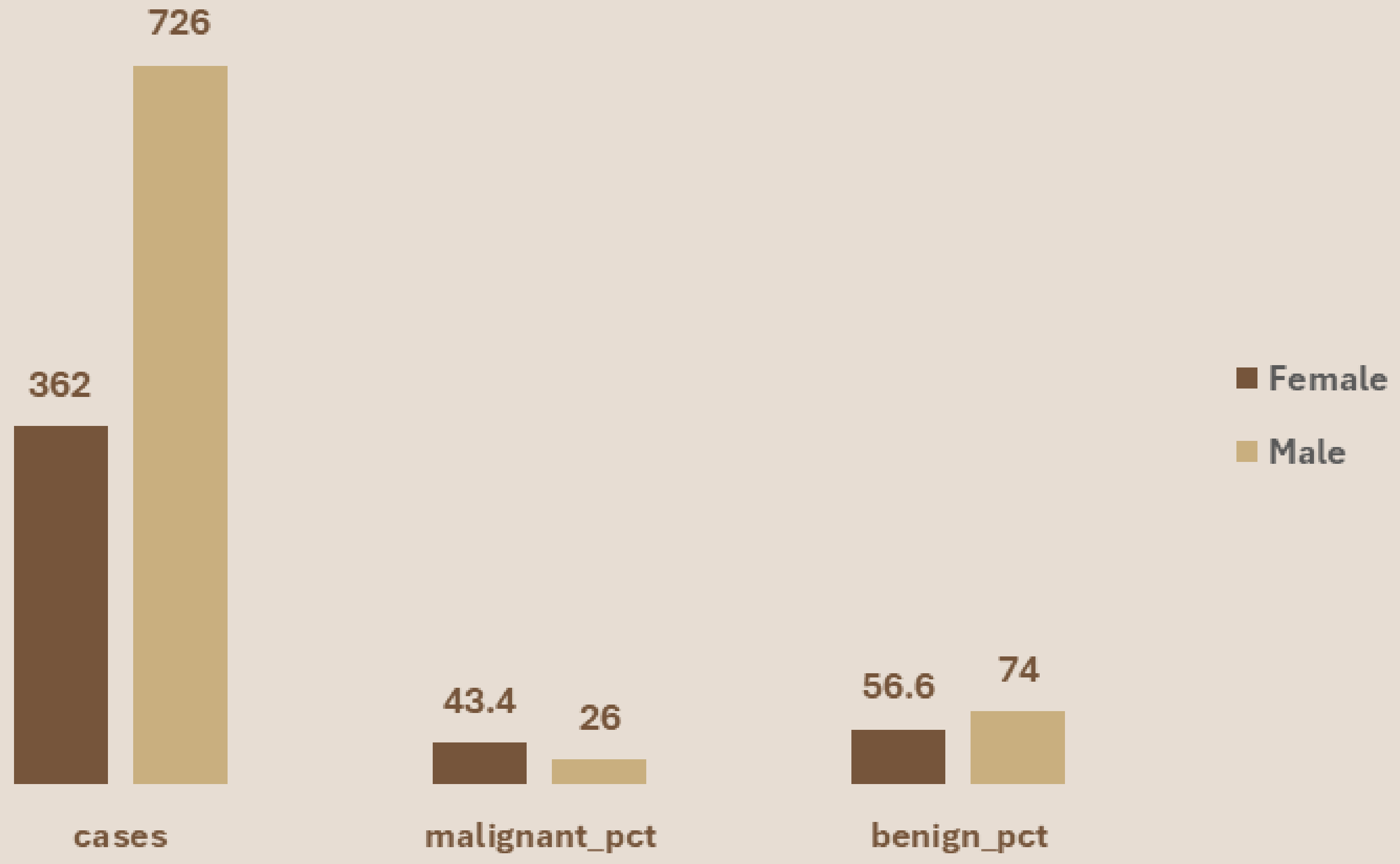
- **Odds of malignancy:** Female odds ≈ 0.77, male odds ≈ 0.35 → ~2.2× higher odds in females.

- **Where malignancies occur:** Of 346 malignant cases, ~45.4% are female (157) and ~54.6% male (189). Females are over-represented in malignancies relative to their case volume (45% of malignants vs 33% of cases).

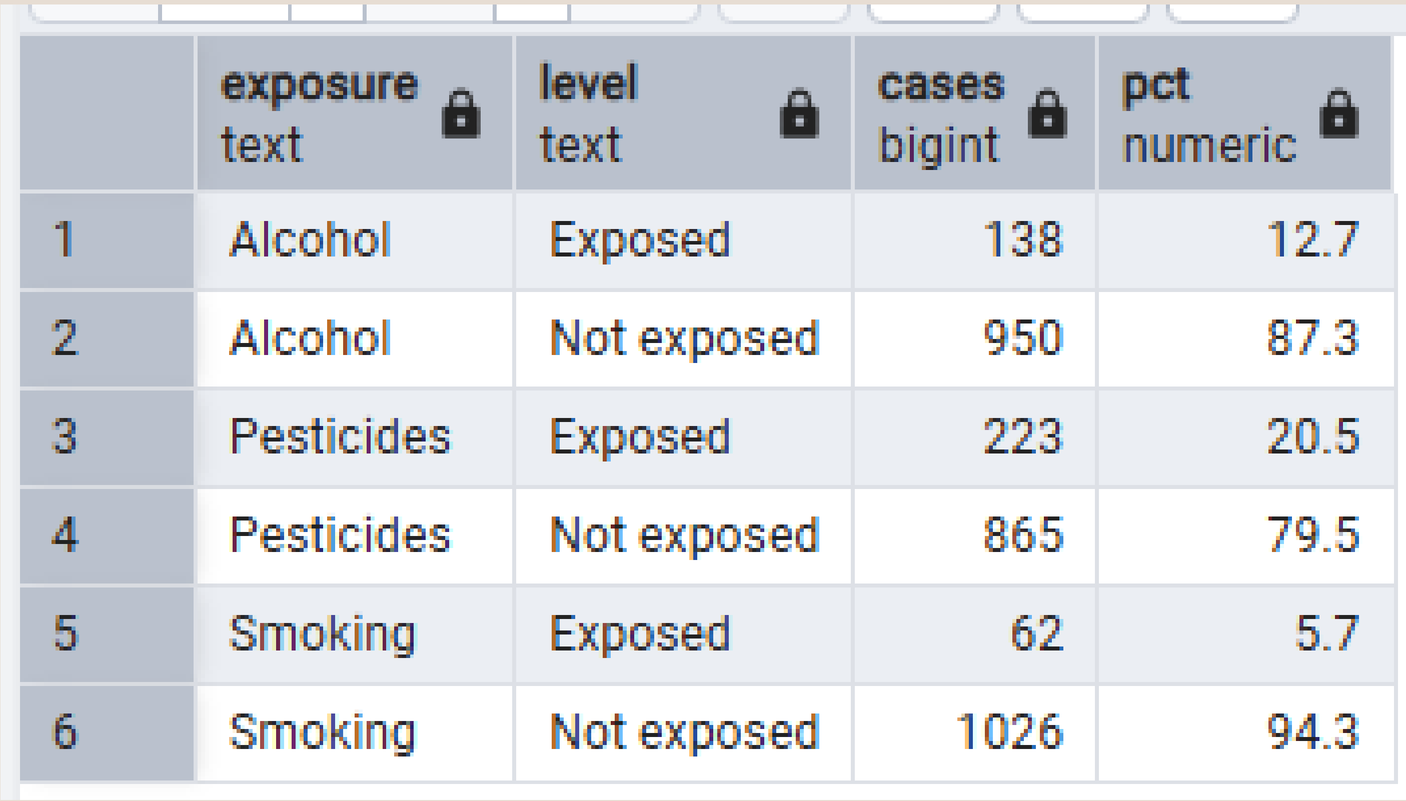
- **Triage efficiency:** Roughly 2.3 female lesions per malignancy vs 3.8 male lesions → prioritizing suspicious female lesions yields more cancers per biopsy.



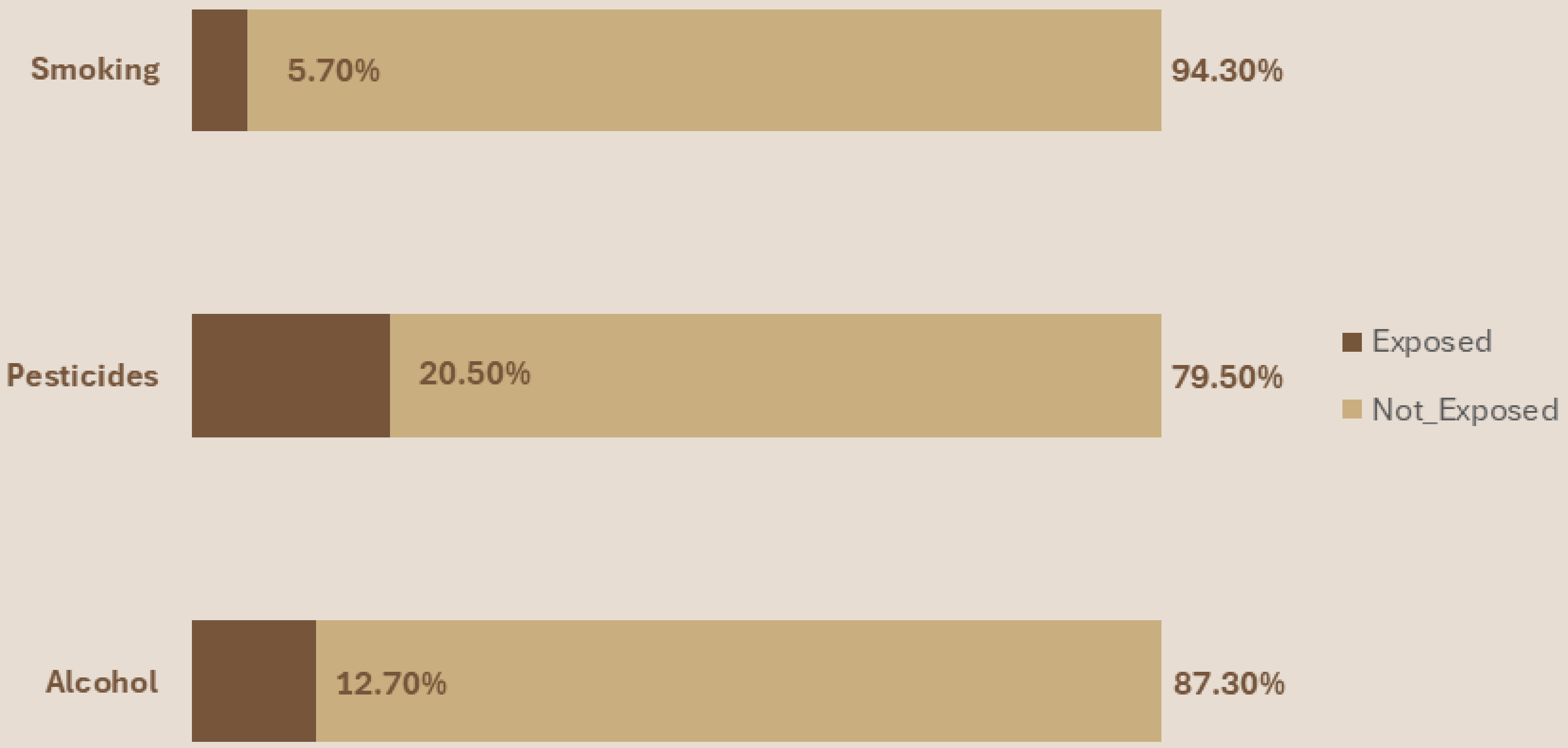
	sex text	cases bigint	malignant_cases bigint	malignant_pct numeric	benign_pct numeric
1	Female	362	157	43.4	56.6
2	Male	726	189	26.0	74.0



Exposure vs Prevalence



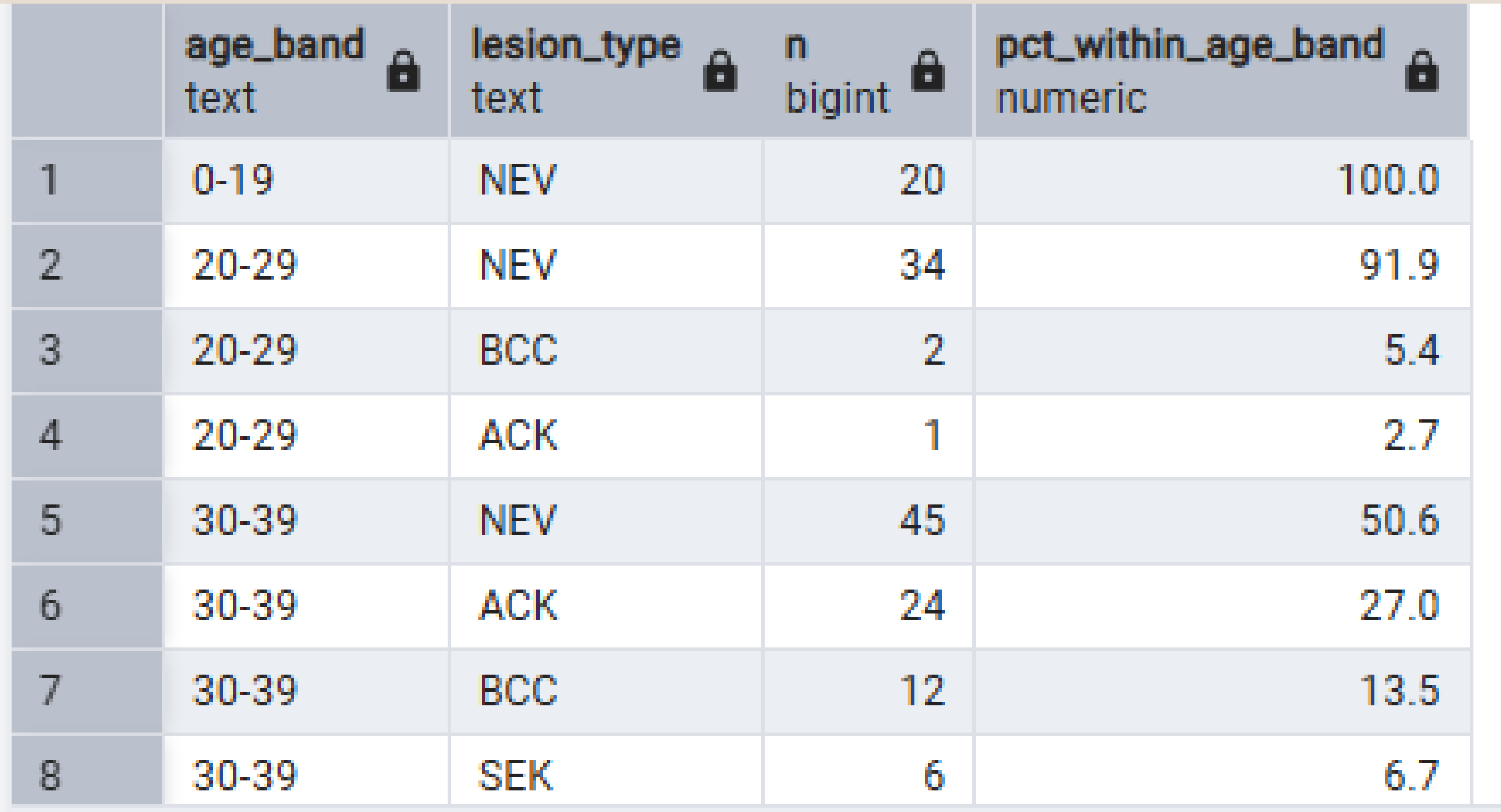
	exposure text	level text	cases bigint	pct numeric
1	Alcohol	Exposed	138	12.7
2	Alcohol	Not exposed	950	87.3
3	Pesticides	Exposed	223	20.5
4	Pesticides	Not exposed	865	79.5
5	Smoking	Exposed	62	5.7
6	Smoking	Not exposed	1026	94.3



- Pesticide exposure is most common: 20.5% (223/1,088).
- Alcohol exposure is moderate: 12.7% (138/1,088).
- Smoking exposure is rare: 5.7% (62/1,088).

Lesion Type by Age Band

- Younger cohorts are overwhelmingly benign NEV (100% at 0–19; ~92% at 20–29), but by 30–39 the benign share halves and BCC emerges (~13%), showing an age-driven shift toward malignancy.

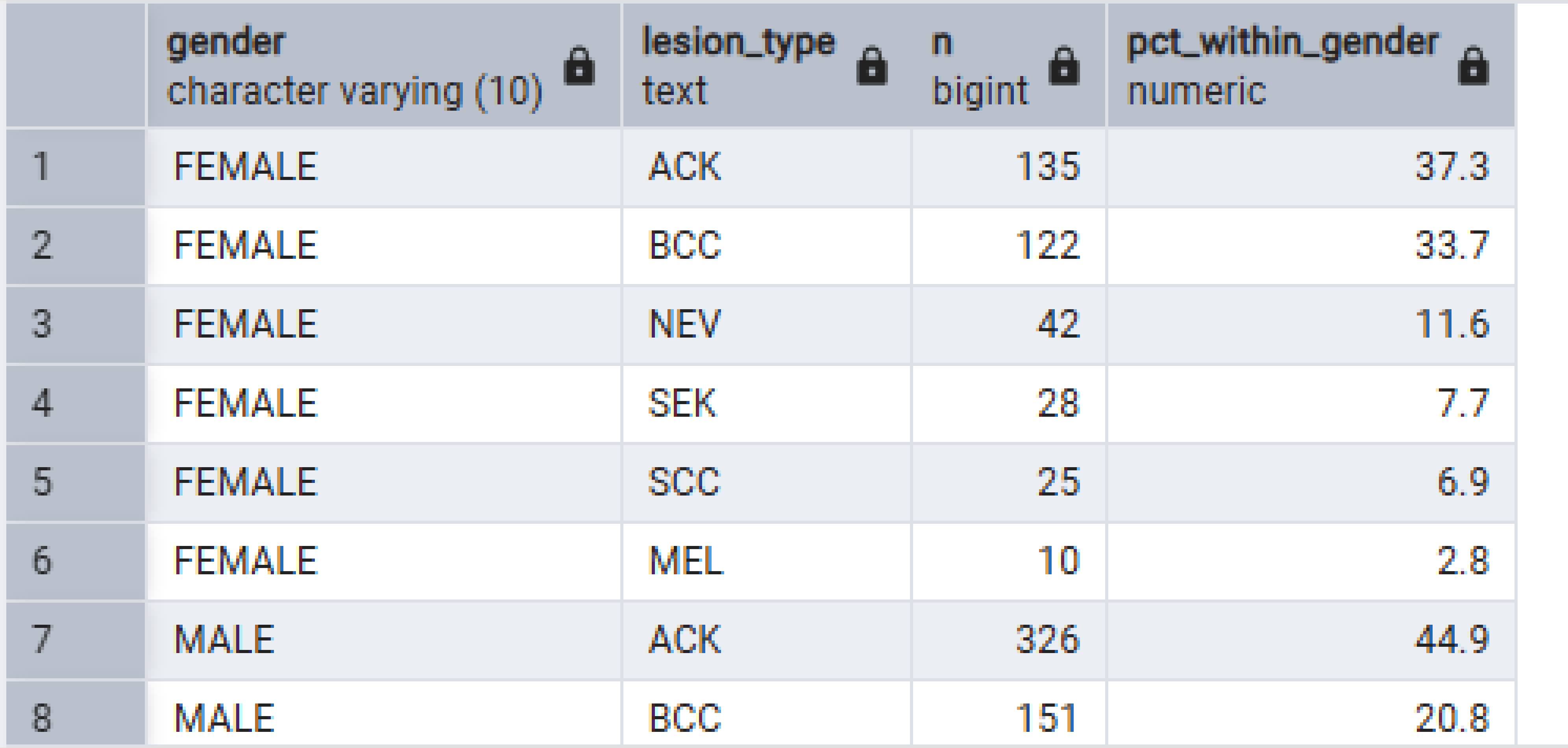


	age_band text	lesion_type text	n bigint	pct_within_age_band numeric
1	0-19	NEV	20	100.0
2	20-29	NEV	34	91.9
3	20-29	BCC	2	5.4
4	20-29	ACK	1	2.7
5	30-39	NEV	45	50.6
6	30-39	ACK	24	27.0
7	30-39	BCC	12	13.5
8	30-39	SEK	6	6.7

Lesion Type by Gender

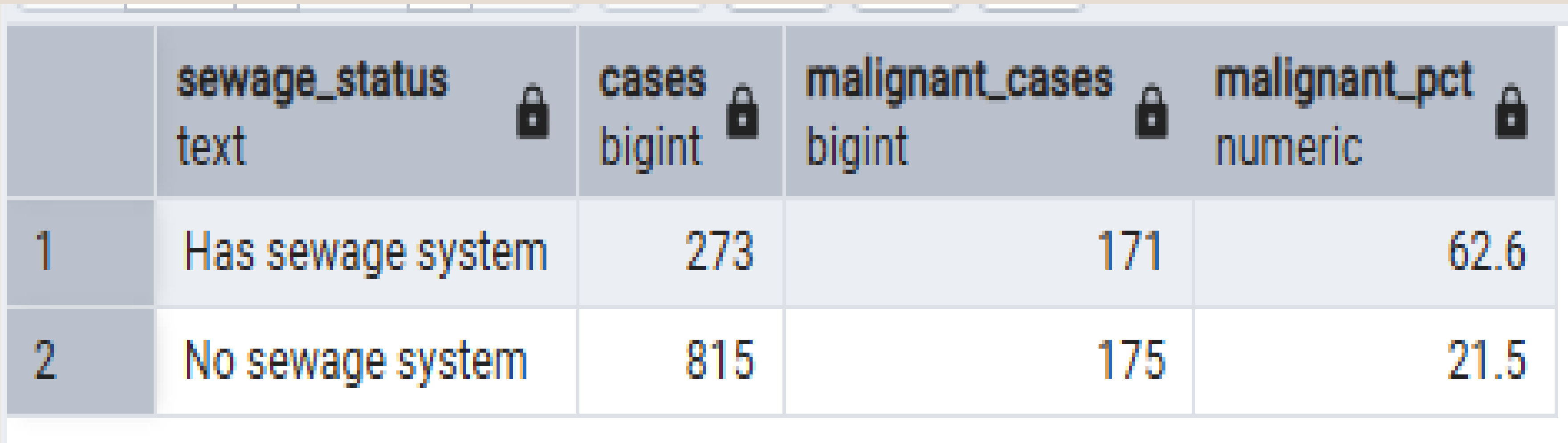
- ➢ Within gender, BCC is the leading cancer subtype forming a larger share of female lesions (~34%) than male (~21%)

- ➢ while males show more actinic keratoses (ACK) (~45%), indicating different lesion profiles by sex.



	gender character varying (10)	lesion_type text	n bigint	pct_within_gender numeric
1	FEMALE	ACK	135	37.3
2	FEMALE	BCC	122	33.7
3	FEMALE	NEV	42	11.6
4	FEMALE	SEK	28	7.7
5	FEMALE	SCC	25	6.9
6	FEMALE	MEL	10	2.8
7	MALE	ACK	326	44.9
8	MALE	BCC	151	20.8

Malignancy vs. Sewage System



	sewage_status text	cases bigint	malignant_cases bigint	malignant_pct numeric
1	Has sewage system	273	171	62.6
2	No sewage system	815	175	21.5

- **Massive risk gap:** Patients with a sewage system have a 62.6% malignancy rate vs 21.5% without—a +41.1 pp uplift.

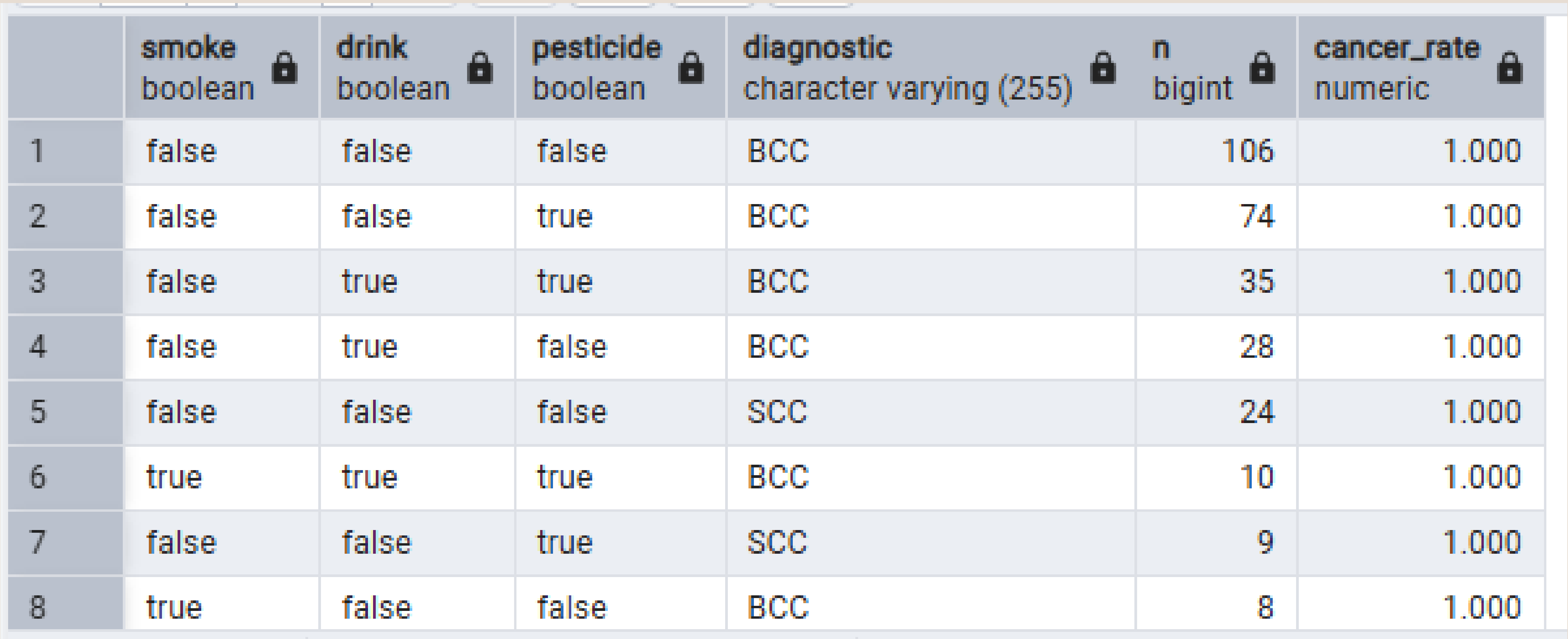
- **~3× higher risk;** ~6× higher odds: Risk ratio ≈ 2.9x (0.626/0.215). Odds ratio ≈ 6.1x.

- **Yield difference:** "Has sewage" needs ~1.6 lesions per cancer (273/171) vs 4.7 without—3× better biopsy yield.

- **Contribution vs volume:** Only 25.1% of lesions are in the "has sewage" group, but they account for 49.4% of all cancers (171/346).

Environmental Factors vs. Lesion type

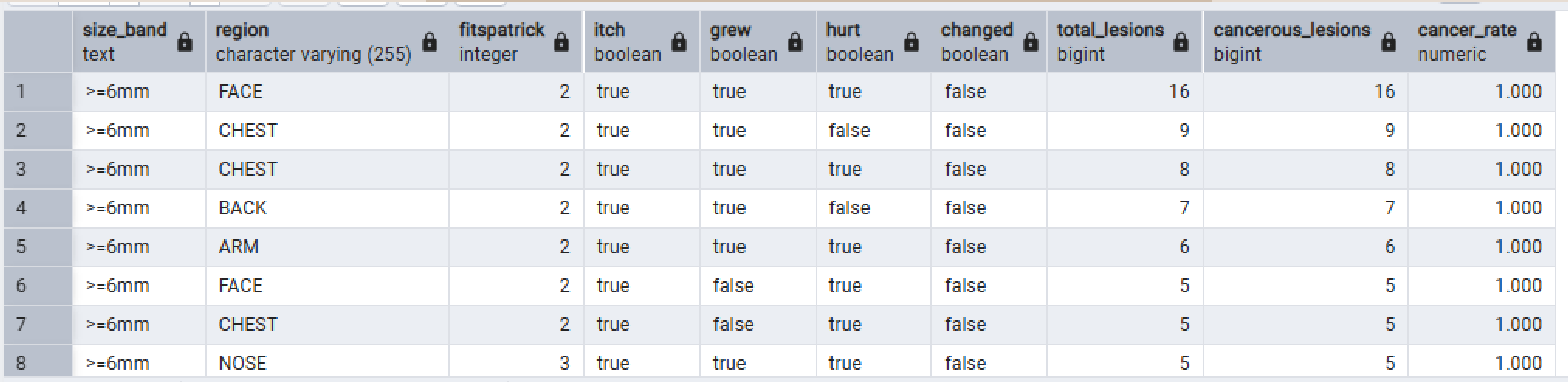
- Among cancerous lesions, BCC dominates across exposure profiles; pesticide exposure appears frequently in BCC, yet the largest cluster is with no exposures, indicating only a modest environmental correlation.



	smoke boolean	drink boolean	pesticide boolean	diagnostic character varying (255)	n bigint	cancer_rate numeric
1	false	false	false	BCC	106	1.000
2	false	false	true	BCC	74	1.000
3	false	true	true	BCC	35	1.000
4	false	true	false	BCC	28	1.000
5	false	false	false	SCC	24	1.000
6	true	true	true	BCC	10	1.000
7	false	false	true	SCC	9	1.000
8	true	false	false	BCC	8	1.000

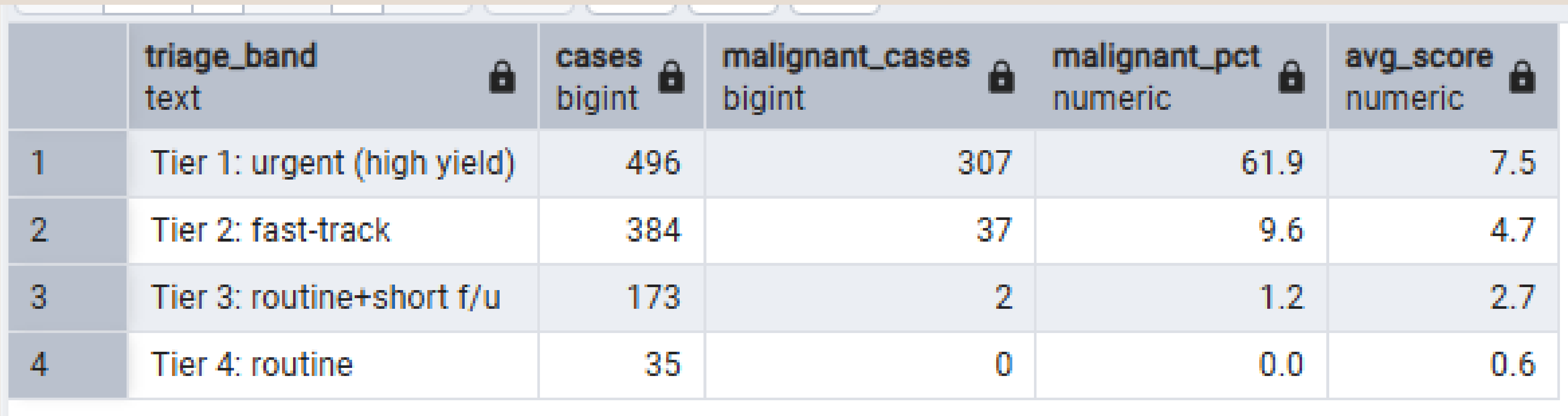
Lesion Characteristics → Cancer vs Benign

- Size ≥6 mm, especially with growth on sun-exposed regions (face/chest/back)—shows ~100% cancer rates, while <6 mm no-growth lesions are overwhelmingly benign.



	size_band text	region character varying (255)	fitspatrick integer	itch boolean	grew boolean	hurt boolean	changed boolean	total_lesions bigint	cancerous_lesions bigint	cancer_rate numeric
1	>=6mm	FACE	2	true	true	true	false	16	16	1.000
2	>=6mm	CHEST	2	true	true	false	false	9	9	1.000
3	>=6mm	CHEST	2	true	true	true	false	8	8	1.000
4	>=6mm	BACK	2	true	true	false	false	7	7	1.000
5	>=6mm	ARM	2	true	true	true	false	6	6	1.000
6	>=6mm	FACE	2	true	false	true	false	5	5	1.000
7	>=6mm	CHEST	2	true	false	true	false	5	5	1.000
8	>=6mm	NOSE	3	true	true	true	false	5	5	1.000

Patterns that support early detection



	triage_band text	cases bigint	malignant_cases bigint	malignant_pct numeric	avg_score numeric
1	Tier 1: urgent (high yield)	496	307	61.9	7.5
2	Tier 2: fast-track	384	37	9.6	4.7
3	Tier 3: routine+short f/u	173	2	1.2	2.7
4	Tier 4: routine	35	0	0.0	0.6

- Tier 1 (urgent) concentrates cancers extremely well: 61.9% malignant (307/496).It’s 45.6% of the workload but captures 88.7% of all cancers (307/346).Lesions per cancer (NNB) ≈ 1.6 → very efficient.

- Tier 2 (fast-track) is low-yield: 9.6% malignant (37/384).Lesions per cancer ≈ 10.4. This looks more like “rule-out” than “fast-track.”

- Tier 3 (routine + short f/u): 1.2% malignant (2/173) → NNB ≈ 86.5.

- Tier 4: 0% malignant (0/35).Overall baseline: 31.8% malignant (346/1088).

Key Insights

- **Age = volume & risk:** 45+ hold 81% of lesions; malignancy rises 3.5% → 20.3% → 33.7% → 36.6%; 90.8% of cancers are in 45+.
- **Sex matters:** Females 43.4% malignant vs males 26.0% → higher biopsy yield per female lesion.
- **Triage works:** Tier 1 = 61.9% malignant, capturing ~89% of cancers with ~46% of workload; Tiers 2–4 are low-yield.
- **Exposures modest:** sewage correlation likely confounded: Pesticides 20.5%, alcohol 12.7%, smoking 5.7%; “has sewage” 62.6% vs 21.5% without—check by age/sex
- **Lesion characteristics** separate malignant from benign.– Size ≥6 mm and recent change/growth/bleeding/pain on sun-exposed regions (face/chest/back) show ~near-certain cancer rates in this set; <6 mm without change is largely benign.

Recommendations

- Prioritize Tier 1 and 45+ (esp. 60+), with extra attention to female patients.
- Auto-flag lesions ≥6 mm or that grew/changed/bleed/hurt/itch → Tier 1.
- Tighten/split Tier 2; route low-signal cases to telederm/routine.
- Next analyses: stratify exposures by age/sex, add simple logistic model, and track NNB & time-to-biopsy by tier.
- Don’t use “has_sewage_system” as a triage rule. Treat it as a proxy (urban/age/access) rather than a causal risk factor. Keep age, sex, lesion size/symptoms as the primary drivers.

Conclusion

- Age and morphology drive risk—apply Tier-1 triage to 45+ (especially women) and changing or ≥6 mm lesions to catch ~9/10 cancers while working up <1/2 of cases.

Thank You