**DERMAI DIAGNOSTICS – SKIN CANCER ANALYSIS  
SQL Project Documentation**

**Table of Contents**

1. [Business Introduction](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#business-introduction)
2. [Problem Statement](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#problem-statement)
3. [Rationale for the Project](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#rationale-for-the-project)
4. [Aim of the Project](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#aim-of-the-project)
5. [Data Sources](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#data-sources)
6. [Project Methodology](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#project-methodology)
7. [Database Setup and Data Exploration](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#database-setup-and-data-exploration)
8. [Data Integration](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#data-integration)
9. [Query Development and Analysis](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#query-development-and-analysis)
10. [Workflow and Outcomes](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#workflow-and-outcomes)
11. [Summary](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#summary)
12. [Recommendations](https://claude.ai/chat/267d7260-aa94-4601-ac5d-78aae65ce1bb#recommendations)

**Business Introduction**

Skin cancer is one of the most common and life-threatening diseases, yet early detection can significantly improve survival rates. Our business, DermAI Diagnostics, leverages machine learning and clinical dermatology research to enhance early diagnosis and treatment. By analyzing patient demographics, environmental factors, and lesion characteristics, we create data-driven insights to assist dermatologists in early-stage detection and decision-making.

Through a combination of AI-powered diagnostic tools, real-world clinical data, and SQL-based research, our company aims to bridge the gap between medical practitioners and machine learning-based skin lesion classification. By providing structured datasets and digital tools, we support medical research, epidemiological studies, and AI-driven skin cancer detection, ultimately improving public health outcomes.

**Problem Statement**

Skin cancer detection is often delayed due to misdiagnosis, lack of access to dermatologists, and limited understanding of environmental risk factors. With 1,089 instances of skin lesions in our dataset, we aim to uncover key patterns linking demographics, environmental exposures, and lesion characteristics to different types of skin cancer. This project seeks to enhance early-stage diagnosis and machine learning-based decision support by structuring data for SQL queries, analysis, and model training.

**Rationale for the Project**

* **Bridging Data and Medicine**: Connecting clinical data with analytical insights to improve diagnostic accuracy
* **Early Detection & Prevention**: Supporting early-stage skin cancer detection through data-driven analysis
* **AI-Driven Medical Research**: Creating structured datasets for machine learning model development
* **Real-World Application**: Providing practical tools for dermatologists and healthcare practitioners
* **SQL Learning Opportunity**: Developing analytical skills while addressing real healthcare challenges

**Aim of the Project**

* Develop a SQL database for students to practice joining clinical and lesion data for effective skin cancer analysis
* Identify environmental and demographic risk factors that correlate with specific skin lesions
* Analyze lesion characteristics to find patterns that indicate cancerous vs. benign lesions
* Create a machine learning-ready dataset to support AI-based early detection of skin cancer
* Enhance dermatological research by providing a well-organized dataset for epidemiological studies and AI model training

**Data Sources**

**Patient Info Table**

**Purpose**: Contains demographic, lifestyle, and environmental information about patients

**Key Fields**:

* 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)

**Data Volume**: Part of 1,089 total lesion instances **Data Quality**: Clean demographic and environmental data with boolean flags for risk factors

**Lesion Info Table**

**Purpose**: Contains detailed information about lesions identified in patients

**Key Fields**:

* 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, ACK, SEK, SCC)
* 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)
* bleeds: 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)

**Data Volume**: 1,089 lesion records **Data Quality**: Comprehensive clinical measurements with symptom indicators and diagnostic classifications

**Project Methodology**

**Technology Stack**

* **Database Management System**: PostgreSQL
* **Interface**: pgAdmin
* **Query Language**: SQL
* **Data Visualization**: Microsoft Excel
* **Presentation**: Microsoft PowerPoint

**Approach**

The project followed a systematic approach to ensure comprehensive analysis and reliable results:

1. **Database Setup and Data Exploration**
2. **Data Integration**
3. **Query Development and Analysis**
4. **Data Visualization and Presentation**
5. **Results Interpretation and Reporting**

**Database Setup and Data Exploration**

**Step 1: Database Creation and Data Import**

**Objective**: Establish a working database environment with the source data

**Process**:

1. **Database Creation**: Created a new PostgreSQL database using pgAdmin
2. **Data Import**: Restored the database by importing the two source tables:
   * Patient Info Table
   * Lesion Info Table
3. **Initial Verification**: Confirmed successful import by checking table structures and record counts

**Tools Used**: PostgreSQL, pgAdmin

**Step 2: Data Exploration**

**Objective**: Understand the structure, quality, and characteristics of the data

**Process**:

1. **Data Quality Assessment**: Evaluated each variable in both tables for:
   * Missing values
   * Data distribution
   * Potential outliers
   * Data consistency issues
2. **Exploratory Queries**: Ran basic queries to understand:
   * Record counts
   * Unique values in key fields
   * Data ranges and patterns

**Key Findings**:

* Dataset contains 1,088 lesion instances across multiple diagnostic categories
* Six main diagnostic types: BCC (Basal Cell Carcinoma), MEL (Melanoma), NEV (Nevus), ACK (Actinic Keratosis), SEK (Seborrheic Keratosis), SCC (Squamous Cell Carcinoma)
* Clean data structure with consistent boolean flags for symptoms and risk factors
* Comprehensive demographic coverage across age groups and genders
* Environmental risk factors well-documented (pesticide exposure, water/sewage access)
* Lesion measurements provided in dual diameter format for area calculations
* No significant missing values identified

**Data Integration**

**Step 3: View Creation for Data Integration**

**Objective**: Combine the two tables into a unified dataset to facilitate analysis and reduce query complexity

**Process**:

1. **Relationship Analysis**: Identified the common fields between Patient Info and Lesion Info tables
2. **View Design**: Created a SQL view that joins the two tables using appropriate join conditions
3. **View Validation**: Tested the view to ensure data integrity and completeness
4. **Performance Optimization**: Ensured the view performs efficiently for subsequent queries

**SQL Implementation**:

-- Table renaming for clarity

ALTER TABLE table1 RENAME TO Patient\_Info;

ALTER TABLE table2 RENAME TO Lesion\_Info;

-- Create integrated view

CREATE VIEW lesion\_patient\_view AS

SELECT

li.\*,

pi.age,

pi.gender,

pi.pesticide,

pi.has\_piped\_water,

pi.has\_sewage\_system,

pi.smoke,

pi.drink,

pi.background\_father,

pi.background\_mother,

pi.skin\_cancer\_history,

pi.cancer\_history

FROM

lesion\_info li

JOIN

patient\_info pi ON li.patient\_id = pi.patient\_id;

**Benefits of Integration**:

* Simplified query writing
* Reduced redundancy in SQL code
* Improved query performance
* Easier maintenance and updates

**Data Visualization and Presentation**

**Step 5: Data Export and Visualization**

**Objective**: Transform SQL query results into meaningful visualizations for stakeholder communication

**Process**:

1. **Data Export**: Exported query results from PostgreSQL to Microsoft Excel using CSV format
2. **Chart Creation**: Developed comprehensive data visualizations.
3. **Visualization Design**: Applied consistent formatting and colour schemes to enhance readability
4. **Presentation Development**: Created PowerPoint presentation incorporating all the visuals

**Visualization Components**:

* **Environmental Risk Analysis Charts**: Visual comparison of pesticide exposure, water access, and sewage system availability across diagnostic categories
* **Demographic Risk Profiles**: visual showing demographic risk factors by diagnostic categories
* **Lesion Size Comparisons**: Treemap chart displaying average lesion areas by diagnostic type
* **Symptom Indicator Matrix**: Comparative visualization of symptom prevalence across different lesion types
* **High-Risk Patient table**: Summary table highlighting key characteristics of identified high-risk patients

**Tools Used**: Microsoft Excel for data visualization, Microsoft PowerPoint for presentation development

**Step 4: Question Generation and Query Development**

**Objective**: Develop analytical questions based on the problem statement and project aims

**Process**:

1. **Question Formulation**: Generated specific analytical questions aligned with:

* Problem statement requirements
* Project objectives
* Business needs

1. **Query Design**: Developed SQL queries to answer each question
2. **Results Export**: Exported query outputs to Microsoft Excel for data visualization and analysis
3. **Documentation**: Documented each query with clear explanations

**Analytical Questions Developed**:

1. **Environmental Risk Factor Analysis**: What environmental factors correlate with different types of skin lesions?
2. **Demographic Risk Profiling**: How do demographic factors (age, gender, lifestyle) relate to lesion malignancy?
3. **Lesion Size Analysis**: What is the relationship between lesion size and diagnosis type?
4. **Symptom Indicator Mapping**: Which symptoms are most indicative of different lesion types?
5. **High-Risk Patient Identification**: Who are the highest-risk patients based on combined risk factors?

**Workflow and Outcomes**

**Query 1: Environmental Risk Factor Analysis**

**Question**: What environmental factors correlate with different types of skin lesions?

**SQL Query**:

SELECT

CASE

WHEN diagnostic IN ('MEL', 'SCC', 'BCC') THEN 'Malignant'

WHEN diagnostic IN ('ACK') THEN 'Precancerous'

WHEN diagnostic IN ('SEK', 'NEV') THEN 'Benign'

ELSE 'Unknown'

END AS diagnosis\_category,

SUM(CASE WHEN pesticide = TRUE THEN 1 ELSE 0 END) AS pesticide\_exposed,

SUM(CASE WHEN has\_piped\_water = False THEN 1 ELSE 0 END) AS lacks\_piped\_water,

SUM(CASE WHEN has\_sewage\_system = False THEN 1 ELSE 0 END) AS lacks\_sewage\_system,

COUNT(\*) AS total\_lesions

FROM lesion\_patient\_view

GROUP BY diagnosis\_category;

**Results**:

* **Malignant Lesions (346 cases)**: Highest pesticide exposure at 45.09%, 43.64% lacked piped water, 50.58% lacked sewage system
* **Precancerous Lesions (461 cases)**: 13.45% exposed to pesticides, 80.48% lacked piped water, 82% lacked sewage system
* **Benign Lesions (281 cases)**: Only 1.78% exposed to pesticides, 92.53% lacked piped water, 93.24% lacked sewage system

**Business Insight**: Strong associations exist between environmental risk factors and lesion types. Malignant lesions show the highest correlation with pesticide exposure, while benign and precancerous lesions are more prevalent in areas lacking basic utilities. This suggests different environmental pathways to cancer development.

**Query 2: Demographic Risk Factor Profile**

**Question**: How do demographic factors (age, gender, lifestyle) relate to lesion malignancy?

**SQL Query**:

SELECT

CASE

WHEN diagnostic IN ('MEL', 'SCC', 'BCC') THEN 'Malignant'

WHEN diagnostic IN ('ACK') THEN 'Precancerous'

WHEN diagnostic IN ('SEK', 'NEV') THEN 'Benign'

ELSE 'Unknown'

END AS diagnosis\_category,

gender,

CASE

WHEN age < 30 THEN 'Below 30'

WHEN age BETWEEN 30 AND 49 THEN '30-49'

WHEN age BETWEEN 50 AND 69 THEN '50-69'

ELSE '70+'

END AS age\_group,

SUM(CASE WHEN smoke = TRUE THEN 1 ELSE 0 END) AS smokes,

SUM(CASE WHEN drink = True THEN 1 ELSE 0 END) AS drinks,

SUM(CASE WHEN skin\_cancer\_history = True THEN 1 ELSE 0 END) AS has\_skin\_cancer\_history,

SUM(CASE WHEN cancer\_history = True THEN 1 ELSE 0 END) AS has\_cancer\_history,

COUNT(\*) AS total\_lesions

FROM lesion\_patient\_view

GROUP BY 1,2,3

ORDER BY 1,2,3;

**Results**:

* **Males with Malignant Lesions**: Highest smoking (up to 23.66%) and drinking rates (up to 59.26%) in older age groups, significant skin cancer history (up to 51.85%)
* **Females with Malignant Lesions**: Lower smoking (up to 7.41%) and drinking rates (up to 9.88%), but high skin cancer history (up to 54.32%)
* **Benign Lesions**: Generally occur in younger patients with healthier lifestyles across both genders

**Business Insight**: Males show significantly higher lifestyle risk factors, while both genders demonstrate increased risk with age and personal cancer history. This suggests gender-specific prevention strategies are needed.

**Query 3: Lesion Size Analysis**

**Question**: What is the relationship between lesion size and diagnosis type?

**SQL Query**:

SELECT

diagnostic,

ROUND(AVG(3.142 \* diameter\_1/2 \* diameter\_2/2)::numeric, 2) AS avg\_area\_mm2

FROM lesion\_patient\_view

GROUP BY diagnostic

ORDER BY avg\_area\_mm2 DESC;

**Results**:

* **MEL (Melanoma)**: Largest average lesion area (197.58 mm²)
* **BCC (Basal Cell Carcinoma)**: Large average lesion area (106.63 mm²)
* **SCC (Squamous Cell Carcinoma)**: Large average lesion area (100.95 mm²)
* **ACK (Actinic Keratosis)**: Moderate size (24.95 mm²)
* **SEK (Seborrheic Keratosis)**: Small average lesion area (5.91 mm²)
* **NEV (Nevus)**: Smallest average lesion area (3.67 mm²)

**Business Insight**: Malignant lesions tend to be significantly larger than benign and precancerous lesions, making lesion size a valuable early warning indicator for prioritizing diagnostic testing.

**Query 4: Cancer Symptom Indicator Analysis**

**Question**: Which symptoms are most indicative of different lesion types?

**SQL Query**:

SELECT diagnostic,

SUM(CASE WHEN itch = TRUE THEN 1 ELSE 0 END) AS lesion\_itching,

SUM(CASE WHEN bleed = TRUE THEN 1 ELSE 0 END) AS lesion\_bleeding,

SUM(CASE WHEN grew = TRUE THEN 1 ELSE 0 END) AS lesion\_grown,

SUM(CASE WHEN hurt = TRUE THEN 1 ELSE 0 END) AS lesion\_hurts,

SUM(CASE WHEN changed = TRUE THEN 1 ELSE 0 END) AS lesion\_changed\_color,

SUM(CASE WHEN elevation = TRUE THEN 1 ELSE 0 END) AS lesion\_elevates,

SUM(CASE WHEN biopsed = TRUE THEN 1 ELSE 0 END) AS biopsy\_confirmed

FROM lesion\_patient\_view

GROUP BY diagnostic;

**Results**:

* **MEL**: Few symptoms, 17 cases confirmed, none unconfirmed
* **SCC**: Moderate symptoms (itching 43, growth 37), 56 cases confirmed
* **BCC**: High symptoms (itching 212), 273 confirmed, 12 unconfirmed
* **ACK**: High symptoms (itching 350), but many unbiopsied lesions (366)

**Business Insight**: Malignant lesions are generally biopsied even with minimal symptoms, while ACK shows high symptom counts but many unbiopsied cases, suggesting potential diagnostic gaps that need addressing.

**Query 5: High-Risk Patient Identification**

**Question**: Who are the highest-risk patients based on combined risk factors?

**SQL Query**:

SELECT

patient\_id,

diagnostic,

fitspatrick,

region,

gender,

age,

background\_father,

background\_mother,

has\_piped\_water,

has\_sewage\_system,

pesticide,

COUNT(\*) AS high\_risk\_lesion\_count

FROM lesion\_patient\_view

WHERE

skin\_cancer\_history = TRUE

AND age > 50

AND pesticide = TRUE

AND diagnostic IN ('BCC', 'MEL', 'SCC')

GROUP BY

patient\_id,

diagnostic,

fitspatrick,

region,

gender,

age,

background\_father,

background\_mother,

has\_piped\_water,

has\_sewage\_system,

pesticide

ORDER BY high\_risk\_lesion\_count DESC;

**Results**: Identified 54 high-risk patients, primarily older adults (>50 years) with personal history of skin cancer and environmental risks. Most high-risk lesions are BCC, with some MEL and SCC cases distributed across Germany, Pomerania, and Italy.

**Business Insight**: High-risk patients can be systematically identified using multiple risk factors, enabling targeted screening and prevention programs for the most vulnerable populations.

**Summary**

**Key Findings**

This analysis of 1,088 skin lesion cases reveals critical patterns in skin cancer risk factors and diagnostic indicators:

**Environmental Risk Patterns**: Strong associations exist between environmental factors and lesion types. Malignant lesions show the highest correlation with pesticide exposure (45.09%), while benign and precancerous lesions are more prevalent in areas lacking basic utilities (80-93% lacking piped water/sewage systems).

**Demographic Risk Stratification**: Males demonstrate significantly higher lifestyle risk factors (smoking up to 23.66%, drinking up to 59.26%) compared to females, particularly in older age groups. Both genders show increased risk with age and personal cancer history.

**Lesion Size as Diagnostic Indicator**: Malignant lesions (MEL: 197.58 mm², BCC: 106.63 mm², SCC: 100.95 mm²) are significantly larger than benign lesions (NEV: 3.67 mm², SEK: 5.91 mm²), confirming lesion size as a valuable early warning indicator.

**Symptom-Based Diagnostic Gaps**: While malignant lesions are consistently biopsied despite minimal symptoms, ACK lesions show high symptom counts but many remain unbiopsied, indicating potential diagnostic gaps in healthcare delivery.

**Technical Achievements**

* Successfully integrated two separate data sources (patient demographics and lesion characteristics) into a unified analytical view
* Developed comprehensive SQL queries addressing all project objectives using advanced aggregation and conditional logic
* Implemented efficient PostgreSQL database design with optimized joins and views
* Created reusable analytical framework for future skin cancer research
* Exported query results to Microsoft Excel for advanced data visualization and chart creation
* Developed comprehensive PowerPoint presentation with professional charts and graphs for stakeholder communication
* Created visual dashboards that effectively communicate complex medical data to both technical and non-technical audiences

**Data Quality Insights**

* Dataset demonstrated high completeness with consistent boolean flags for symptoms and risk factors
* Clean demographic coverage across age groups and genders enabled robust stratified analysis
* Comprehensive diagnostic classifications (BCC, MEL, SCC, ACK, SEK, NEV) provided detailed lesion categorization
* Dual diameter measurements allowed for accurate lesion area calculations

**Performance Considerations**

* View-based integration eliminated query redundancy while maintaining data integrity
* Efficient grouping and aggregation queries handled 1,089 records with optimal performance
* Conditional logic in CASE statements enabled flexible diagnostic categorization

**Recommendations**

**Environmental Health Interventions**:

* Target public health interventions to improve access to clean water and sewage systems to reduce benign and precancerous lesions
* Implement stricter safety measures to reduce pesticide exposure, given its strong association with malignant lesions

**Clinical Practice Improvements**:

* Prioritize large lesions (>100 mm²) for immediate diagnostic testing and biopsy
* Develop symptom-based triage protocols to prioritize multi-symptom lesions for biopsy
* Address diagnostic gaps in ACK lesions by establishing clear biopsy criteria for high-symptom cases

**Risk-Stratified Screening Programs**:

* Focus screening and prevention efforts on older adults (>50 years) with high-risk profiles
* Implement gender-specific prevention strategies: intensive smoking/drinking cessation programs for males, comprehensive cancer screening for both genders

**Healthcare System Optimization**:

* Establish high-risk patient registries for systematic follow-up of the 54 identified high-risk patients
* Create standardized follow-up protocols for symptom-negative lesions to monitor changes over time

**Technical Recommendations**

**Data Infrastructure**:

* Expand the current SQL framework to include longitudinal patient tracking capabilities
* Implement automated risk scoring algorithms based on the identified demographic and environmental factors
* Develop integration capabilities with AI-powered diagnostic imaging systems

**Research Enhancement**:

* Use the structured dataset for machine learning model development to improve diagnostic accuracy
* Establish data collection protocols for ongoing epidemiological studies
* Create data sharing frameworks for collaborative research initiatives

**Data Management Recommendations**

**Quality Assurance**:

* Implement regular data validation checks for new patient and lesion records
* Establish standardized data entry protocols to maintain consistency across diagnostic categories
* Create automated reporting systems for real-time monitoring of risk factor trends

**Visualization and Reporting**:

* Develop automated dashboard systems that can directly connect to PostgreSQL for real-time chart updates
* Create standardized Excel templates for consistent visualization across different analyses
* Implement interactive PowerPoint templates for regular stakeholder presentations
* Establish data visualization best practices for medical data presentation

**Patient Education and Engagement**:

* Develop educational materials about self-monitoring lesion size and growth patterns
* Create awareness campaigns about environmental risk factors, particularly pesticide exposure
* Promote regular dermatological check-ups for individuals with personal or family cancer history
* Use visual aids from the PowerPoint presentation for patient education sessions