Data exploration and group findings

Group H

Participants of the group: Full name / github name

- Edvinas Gross — edgrobit
- Eduardo Almeida — EduardoGralmeida
- Utku Yoztyurk — UtkuYoztyurk
- Ofri Kela — ofriki
- Ada Camille Bertelsen (missing/ did not participate in the group work)

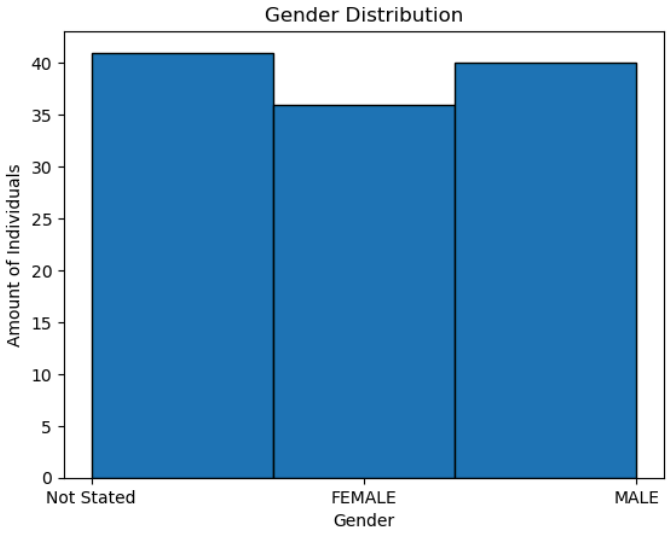
Github repository link:

https://github.com/edgrobit/2026-PDS-GroupH

1.1 Data Set Overview

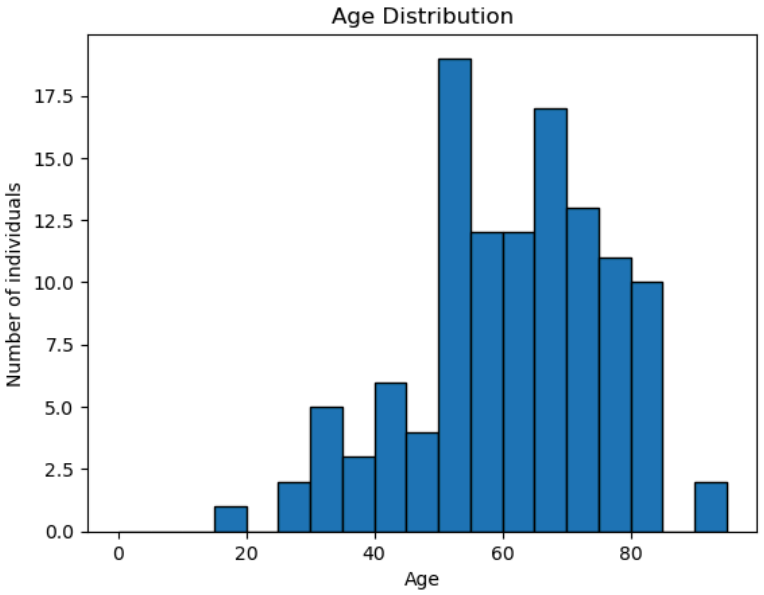
We were initially given a large data set , of which we were only allotted a small partition of it to be explored by us.

To familiarize ourselves with the data, we did a rudimentary scan of our share of the data set, which ended up being 117 entries in our case.



We examined the Gender distribution to understand whether a certain gender was more represented. which were roughly the same, assuming the not-stated entries followed the same trends.

Following this next step was to assess the age distribution of this group.



1.2 Diagnostic Categories

During further exploration of the data, we put our focus as understanding the diagnostic column and what we could learn from it. There were 6 different diagnoses prominent in our data set:

ACK - Actinic Keratosis : a precancerous skin condition due to chronic sun exposure, while it is not technically cancerous it is an alarming condition

NEV - Melanocytic Nevus : A benign mole. While typically harmless, some nevi can develop into melanoma. No treatment is required though monitoring is highly advised.

BCC - Basal Cell Carcinoma : most common type of skin cancer. Slow growing and can cause significant local tissue destruction if not treated.

SEK - Seborrheic Keratosis : A common, benign skin growth. Not cancerous and does not progress into skin cancer though can be mistaken for melanoma.

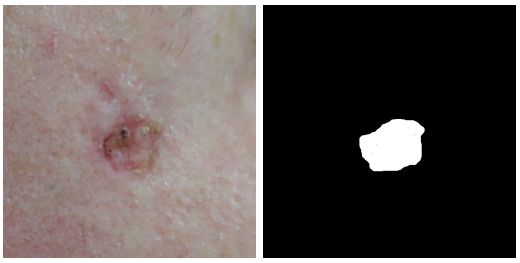
SCC - Squamous Cell Carcinoma : A malignant skin cancer. It often develops from ACK.

MEL - malignant Melanoma : Most aggressive form of skin cancer. Responsible for the majority of skin cancer-related deaths. Early detection is critical.

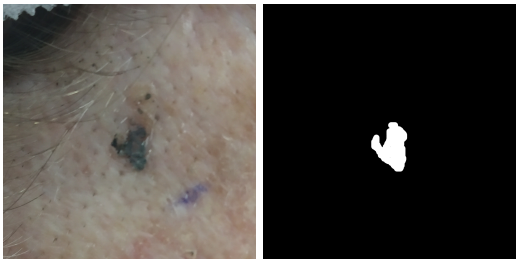
Imagery example of these conditions from the data set:

* a case of Actinic Keratosis (ACK) (PAT_857_1628_916.png)

* a case of Melanocytic Nevus (NEV) (PAT_759_1433_973.png)

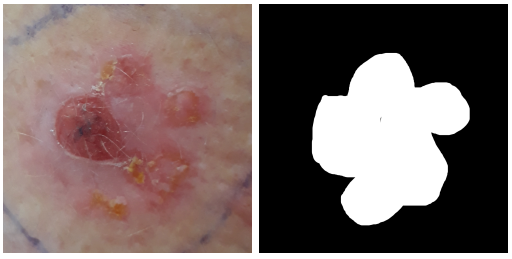


* a case of Basal Cell Carcinoma (BCC) (PAT_944_1795_371.png)



* a case of Seborrheic Keratosis (SEK) (PAT_939_1791_329.png)

* a case of malignant Melanoma (MEL) (PAT_884_1683_538.png)



* a case of Squamous Cell Carcinoma (SCC) (PAT_56_86_802.png)

We decided to focus on the malignant forms of these diagnoses, of them being **BCC**, **SCC** and **MEL.**

1.3 Data Completeness and Case Selection

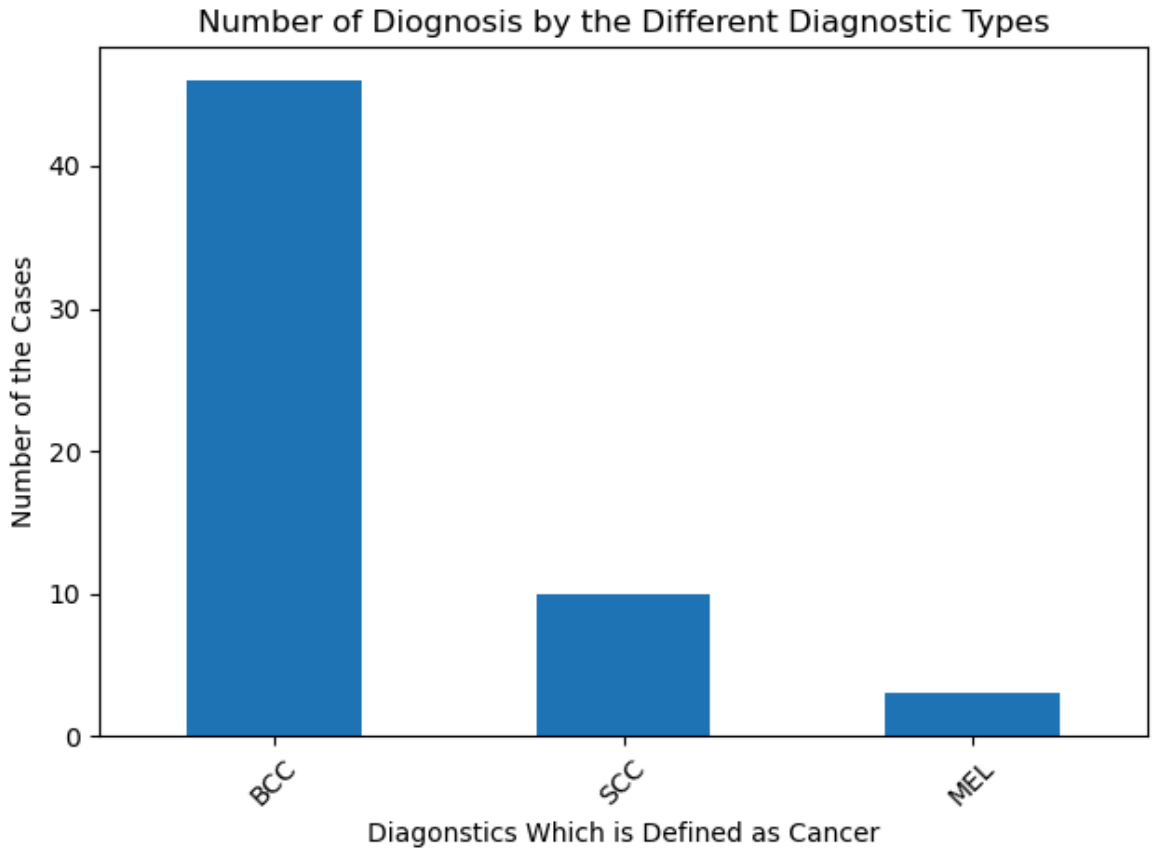
An observation was made that, some of entries were missing a lot of data regarding their assessment, but all of individuals who were diagnosed with malignant formations or were at risk, had all of their data filled out, so an assumption could be made that individuals who were missing data, were of low or no risk so no further intervention was needed. With that assumption some further data analysis was made on individuals with all their data intact as they were most likely to possess our focused points - **BCC**, **SCC** and **MEL.**

2 Data Analysis

In regards to the number of cases themselves 59 of the 117 were diagnosed with cancer, representing roughly a little more than 50% of the entirety of samples.

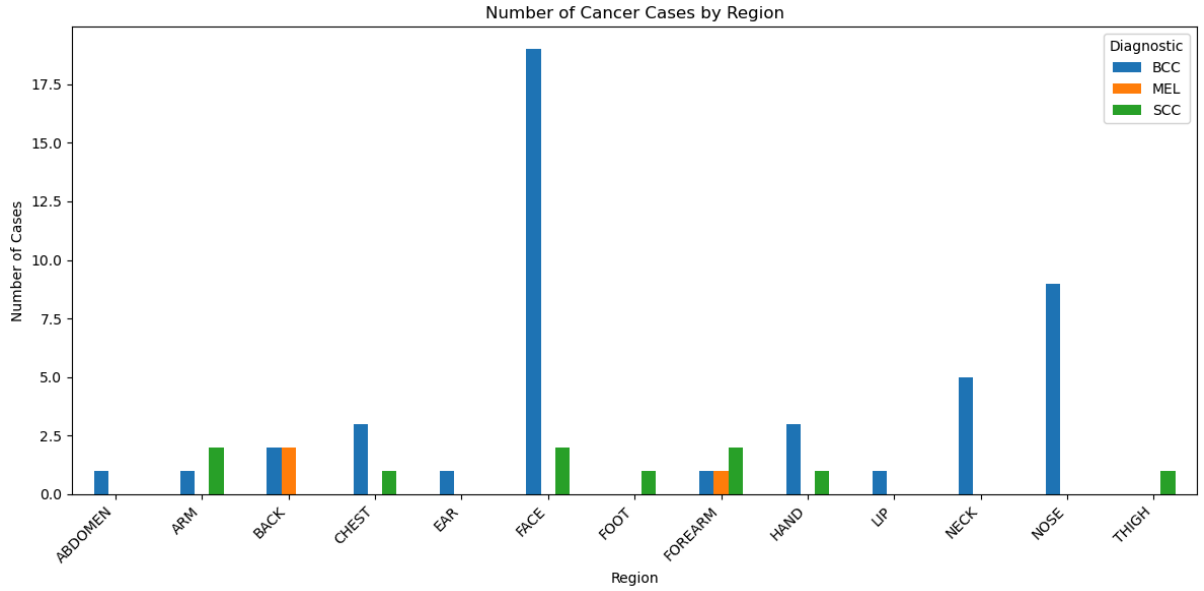
2.1 distribution of cancer cases

Among cancer diagnostics the most common is BCC, with more than 40 cases being observed. We also have very few numbers of MEL, which may suggest that this dataset is not evenly distributed across all cancer types and could be skewed toward certain diagnoses.



2.2 Region of concentration of cancer

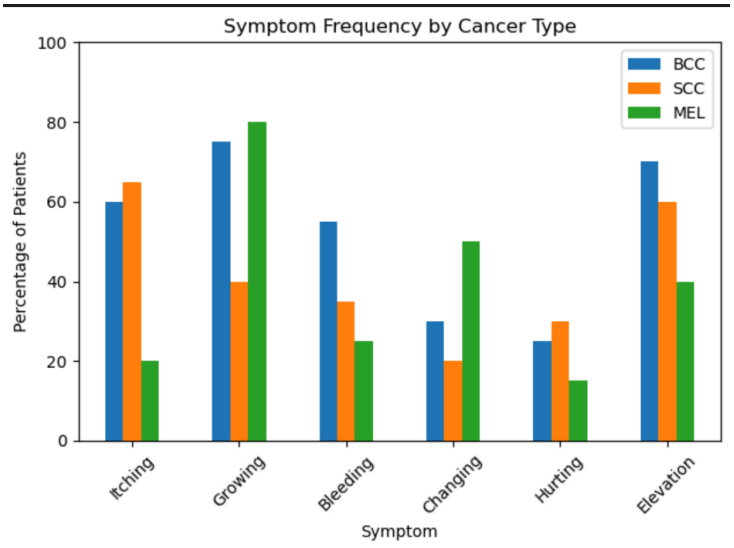
When we look at the region concentration of the cancers, we can see that BCC is highly concentrated on the upper body, in particular on the areas of the face, neck and nose. While there are more occurrences of SCC on the face, arms and forearms and MEL seems to appear only on backs and forearms in our specific dataset.



2.3 Symptom Patterns in Malignant Lesions

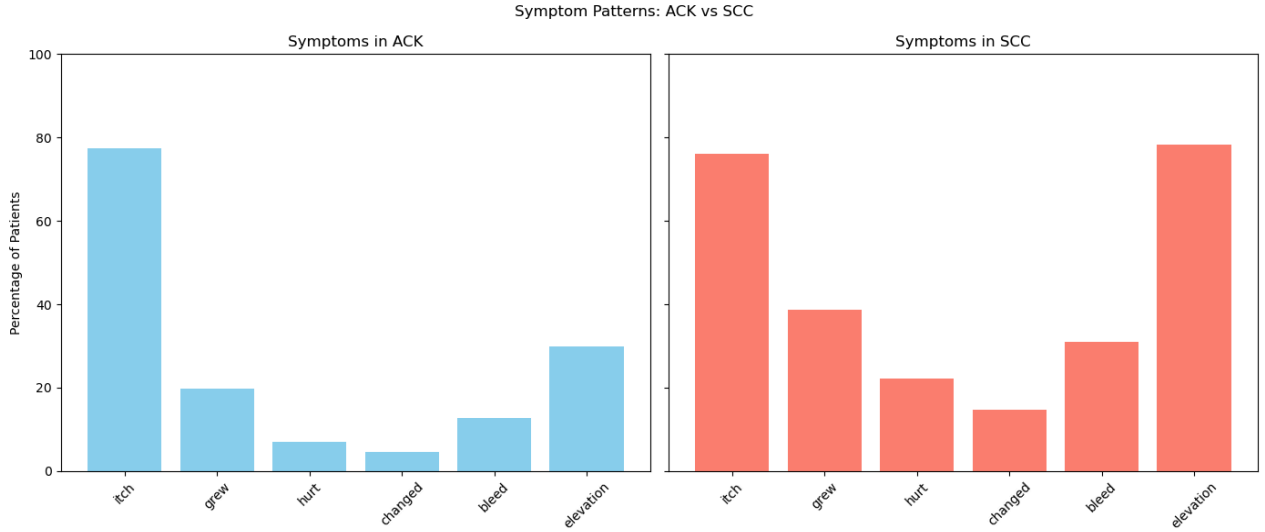
We then looked at each type of malignant lesion, if we could see any pattern into events that occurred with it: itching, growing, hurting, changing, bleeding and if they had any elevation.

We found that for our dataset, the BCCs tend to itch, grow, bleed and have elevation more than not. SCCs tend to itch and have elevation, but they don't change much. MELs though tend to present growth.

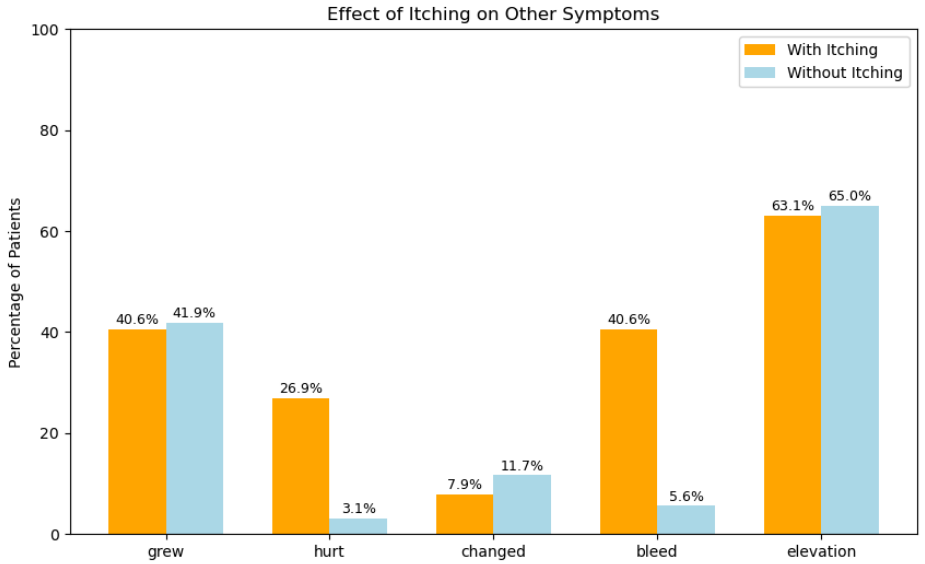


2.4 Symptom Progression from ACK to SCC

From some research on the diseases we found that SCC can develop from ACK. Therefore, we examined whether changes in symptom presentation could be observed between these two diagnoses.
We were also interested in observing the relationship between the symptoms of ACK and SCC, a disease that often develops from ACK. The graph below highlights the difference in the symptoms between these two diagnoses.



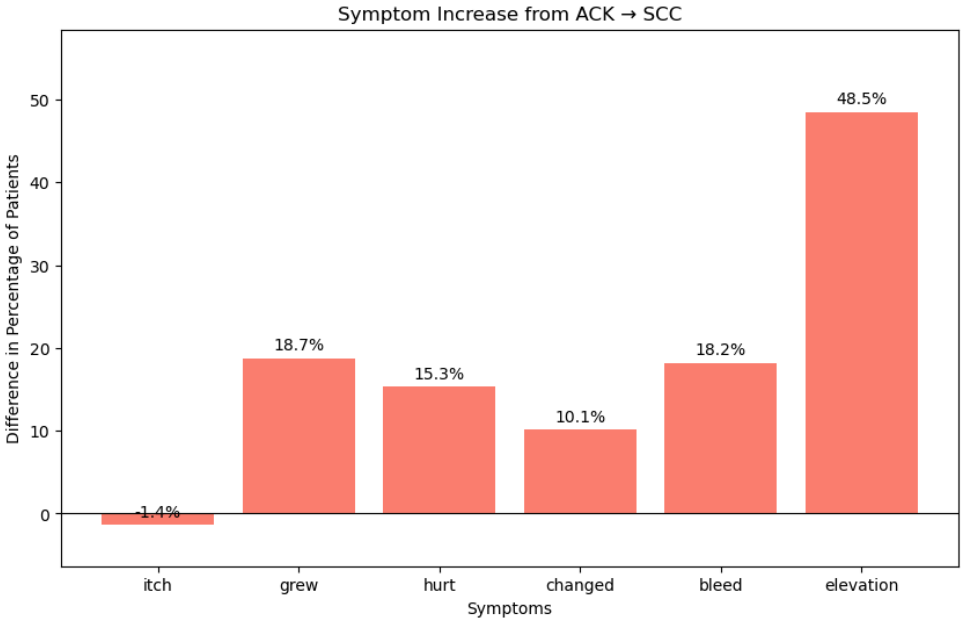
The graphs above suggest that patients diagnosed with ACK primarily report itching as the main symptom, while more severe symptoms are less common. Since itching appears with the highest frequency, we will consider it as the dominant symptom of ACK. Furthermore, it is possible that persistent itching could contribute to the development of other symptoms, such as irritation or visible changes.
The graph below may help with this limitation or the possibility that arises, this information can be checked from the data.



From the data, we observe that symptoms showing the most notable increase in the presence of itching are pain(hurt) and bleeding.
However, when comparing ACK and SCC overall, itching itself is not significantly increased, while other symptoms become more aggressive in SCC.

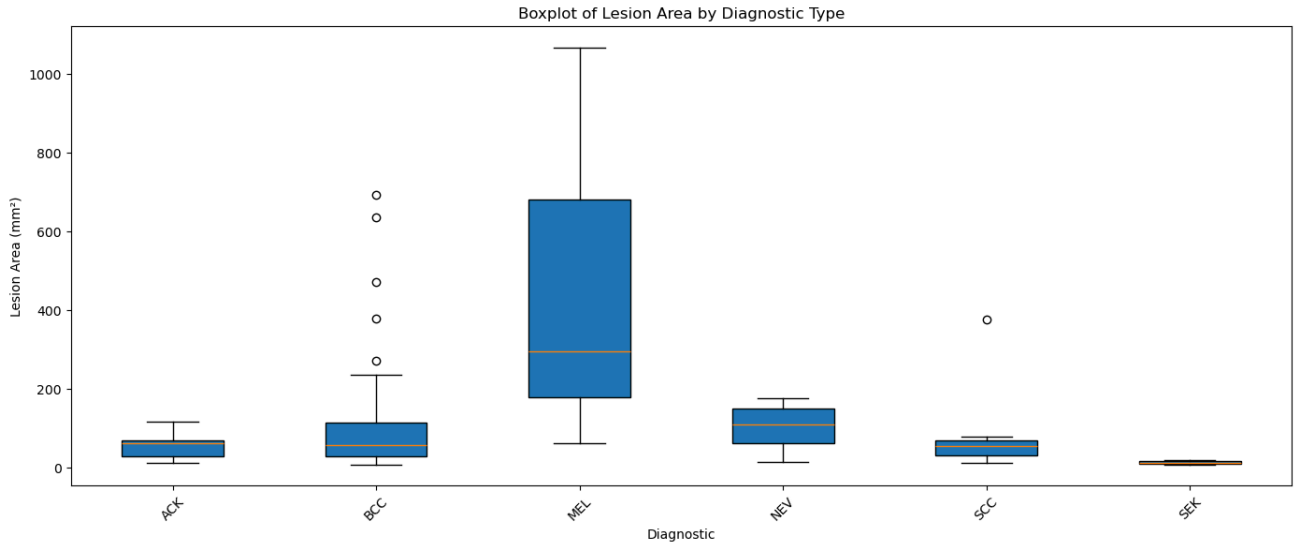
Therefore, although itching may be linked to certain symptoms such as pain and bleeding, it does not seem to be the main factor in the progression of the disease to SCC. Instead, the overall development from ACK to SCC seems to be associated with the increase in more severe and structural symptoms.

This pattern becomes clearer when examining the overall change in symptoms between the two diagnoses.



2.5 Lesion Area Analysis and Diagnostic Implications

These areas were calculated utilizing the diameters of the metadata, and assuming ellipses as the shape of the lesions.

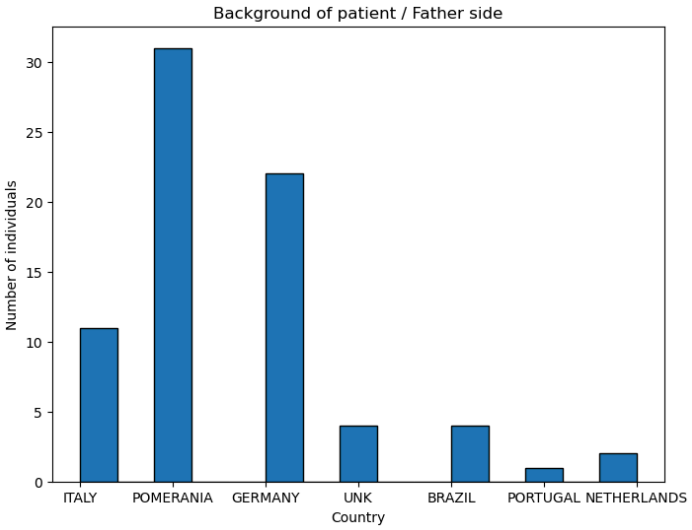


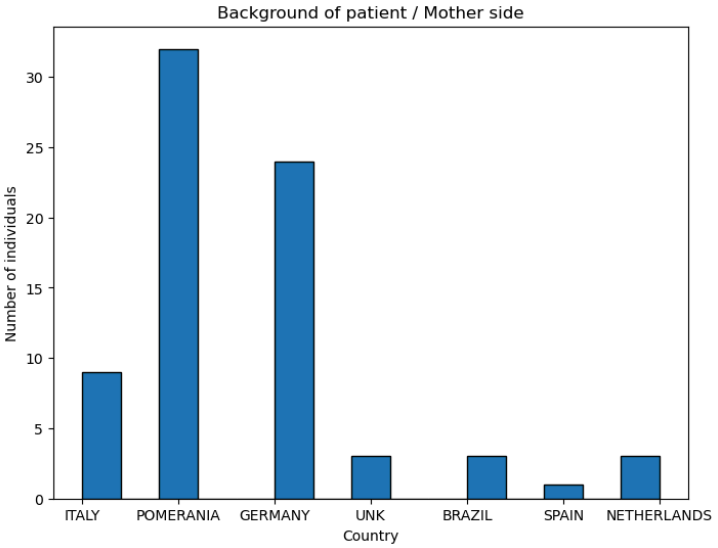
We observe from the median line within each boxplot that the median lesion area is higher for cancerous lesions compared to non-cancerous ones. Indicating that cancer lesions are generally larger. Additionally, the cancer group shows a wider spread and higher extreme values, suggesting that cancer lesions can grow substantially larger. However, lesion size alone does not guarantee malignancy, as some non-cancerous lesions are relatively large and some cancerous lesions are small.

Among the cancer types, MEL demonstrates both the largest average area and the greatest variability, as indicated by the wide interquartile range and extended whiskers. This suggests that melanoma lesions can vary considerably in size, potentially reflecting differences in stage at the time of detection. Therefore, while lesion size appears to be an important supporting factor in identifying cancerous lesions, it should be interpreted alongside other clinical characteristics.

2.6 Ancestry

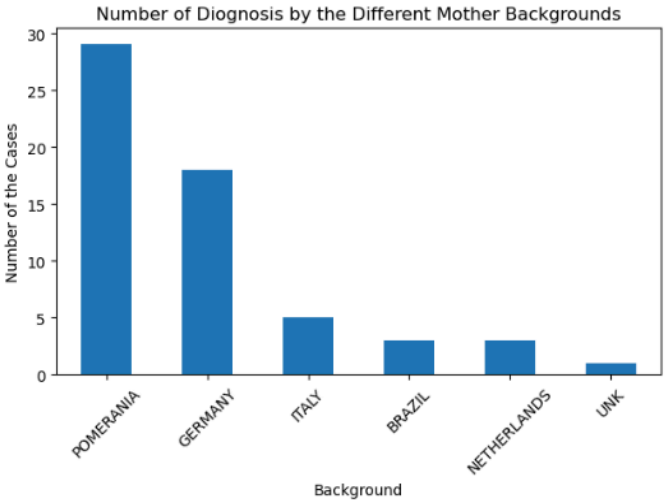
Next step to the assessment was made in regards to ancestry of the individuals at high risk of target conditions.

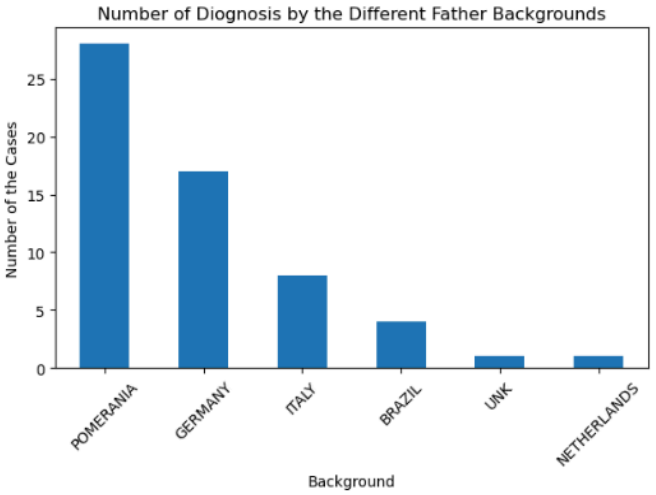




If we dig into the dataset deeper, specifically the cases with the diagnosis “BCC”, “MEL” and “SCC” , we can see that having a background from places such as “Germany” or “Pomerania” (an area split between today's Germany and Poland) actually links with those types of skin cancer patients. While the graphs above showed us where people came from in the dataset, these 2 graphs below will show us the connections with the areas and their connection to those 3 specific skin cancer diagnoses.

As shown on the new graphs, Pomerania is the area with the most cases of those 3 specific



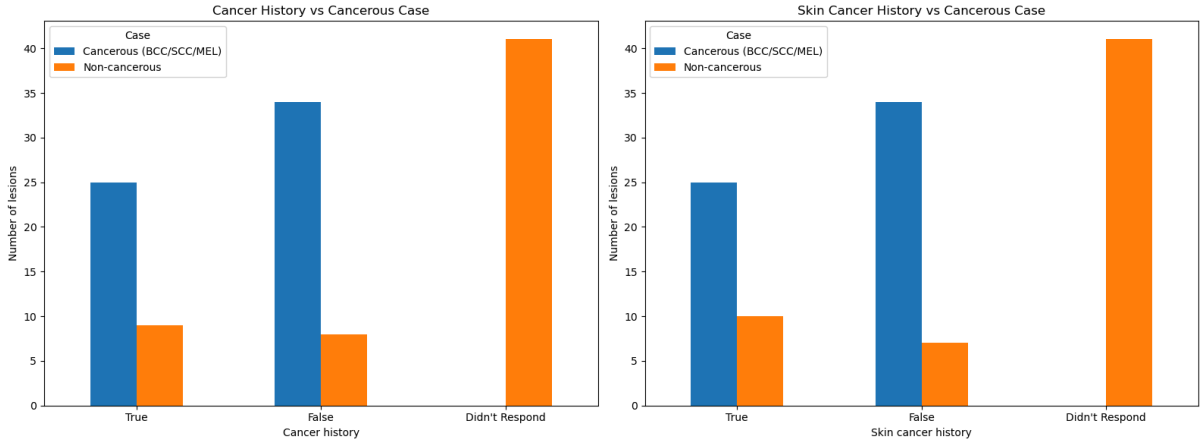


skin cancer diagnoses among the cases in our dataset. Several factors can explain why this specific area shows the highest numbers, such as;

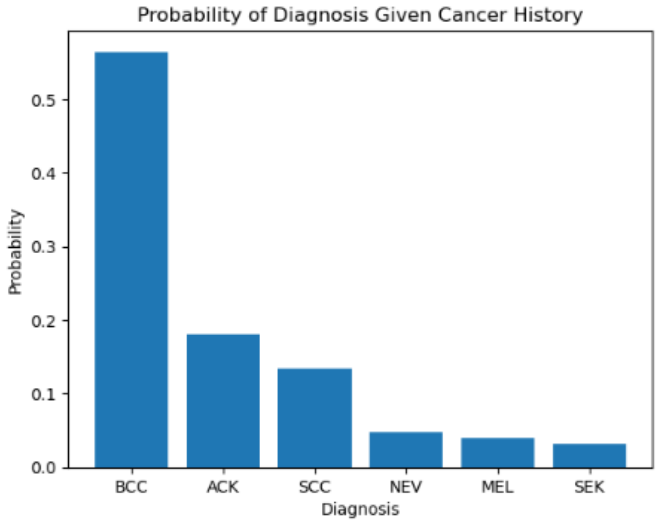
- **Phenotypic Risk Factors:** People from the Pomerania region, have specific features like Fair Skin(Skin Type I or II),light hair and light coloured eyes. These features

have a decreased chance to produce melanin,which increases the chance of getting UV damage.

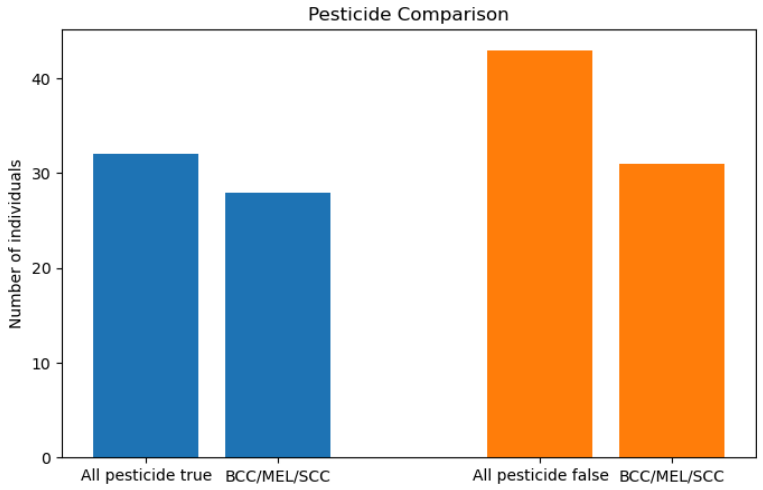
- **Genetic Clusters:** Certain hereditary gene mutations which can increase the risk of getting melanoma and and the other type of skin cancer types can be more prevalent in specific European populations such as Germans and Pomeranian people due to historical reasons or genetical clusters build up in generations. The graph will show the chance of diagnosis with a Cancer background.



According to the data, having cancer or skin cancer history does not appear to be strongly linked with getting a new cancer diagnosis, if anything having a cancer/skin cancer history seems to slightly diminish the chance, maybe because people tend to take more care of themselves.

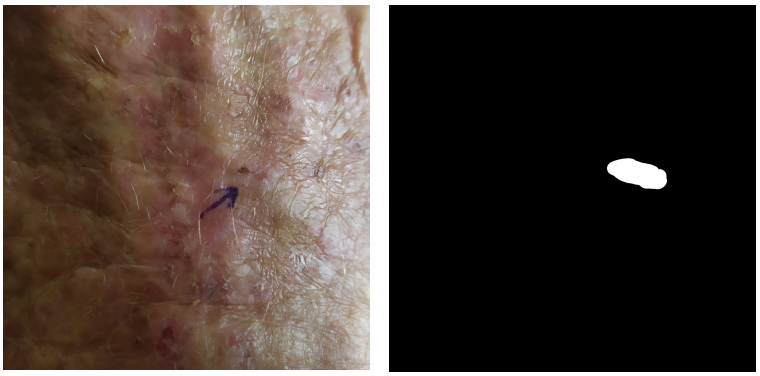


Pesticide exposure was considered as a potential lead of malignant growth appearance and development. We checked how many individuals were exposed to pesticides and were diagnosed positively to individuals who were not exposed to the chemical substances and still had been diagnosed positively.



While the number of diagnoses are relatively the same, the sample size on the pesticides exposed side is considerably smaller, so proportionally the amount of cases is significantly higher. Many concussions could be drawn from this, ranging from can these chemicals be causing these cancerous lesions, or could there be some more complex reasoning, such as people who are exposed to these chemicals tend to work outside and thus have more exposure to the sun. The amount of data to prove or disprove any of these hypotheses is insufficient but worth keeping in mind.

Having done the data exploration, we were tasked to create additional annotations on the data provided. Having to work with an algorithm later on, it was important to create a chart that describes if the mole has been marked with a pen and how much hair is visible in the image. These annotations were crucial later on at making sure the masks were as true to the image as possible and no obstructions were interfering with it. The hair level is marked from a scale of 0-3, where we agreed, that no presence of hair is a 0, but even the slightest amount of hair is a 1, whereas type 2-3 became more variable given the raters judgment. Annotation regarding pen was more consistent though, as no presence meant 0 and 1 was allotted to any markings present.



Examples of annotation:

** presence of pen that would be marked as 1 (PAT_2020_4174_799.png)*

** no presence of hair that would be marked as 0 on hair (PAT_326_690_797.png)*

3 Summary

In this project, we explored 117 skin lesion cases to identify patterns using graphs and data analysis. We examined diagnosis types, symptoms, lesion size, ancestry, age distribution, cancer history, and pesticide exposure to better understand possible risk factors.

Our graphs showed that cancer lesions are usually larger than non-cancer lesions and vary more in size, although size alone does not determine whether a lesion is cancerous. We also observed differences in symptom patterns, and the comparison between ACK and SCC suggested that progression is mainly linked to increased structural symptoms such as growth and elevation.

When analyzing pesticide exposure, we noticed that the exposed group was smaller, but proportionally showed more cancer cases. However, the dataset was not large enough to determine whether pesticides directly cause cancer. Similarly, while ancestry, age, and cancer history were explored, no strong causal conclusions could be drawn.

Throughout the project, we relied on visualizations and careful data exploration to guide our conclusions, while keeping in mind the limitations of the dataset. Additional image annotations were also created to improve data quality for future algorithm development.