# **Design process Report**

## ERD Diagram

After noun analysis I chose 6 tables:

1. Patient
2. Doctor
3. Surgeon
4. Department
5. Referral
6. Waitlist

Then I came to a few assumptions when creating the ERD regarding the relationships between tables.

* The patient can have many doctors and the doctor can have many patients (M:M)
* The Patient has one doctor and doctors can have many patients (M:1)
* Department has many surgeons and surgeons have only one department (1:M)
* Surgeons have many patients, but patients have only one surgeon (1:M)
* Referral has many patients and patients can have many referrals (M:M)
* Referral has many patients and patients can have only one referral (1:M)
* Department has many referrals and referrals can be for many departments (M:M)
* Waitlists can be for many patients and patients can have many waitlists (M:M)
* Waitlists can be for many patients and patients can be on only one waitlist (1:M)
* Department has many waitlists and waitlists can be for many departments (M:M)
* Department can have many patients, but patients can have only 1 department(1:M)

The assumptions that I made, I concluded that red is incorrect.

1. I feel for this instance the patient can have only 1 doctor as the data only shows one doctor.
2. Patients are not linked directly to the surgeon rather they are linked to the department and the surgeons are also linked to the department.
3. For this instance, I feel that patients can have only one referral.
4. The same as above the patients can be on only one waitlist.
5. This was the hardest one to assume but because department has a relationship with referral and waitlist, and referral and waitlist have a relationship with patient then this is the relationship that the department has with patient, it is purely through the referral and waitlist.

**Bridging tables:**

There needs to be two bridging tables one between department and referral and one between department and waitlist to fix the many to many relationships.

So, I concluded with the ERD there needs to be eight tables in total.

## Data Cleansing

1. I deleted the column with no column heading because it didn’t seem to have any significance with the rest of the data.
2. I deleted the column year-month because this can be generated with a query.
3. I formatted all the date fields to match SQL date, that being yyyy-mm-dd
4. I separated the date columns and the string columns so I could trim all the string columns to check for leading and trailing spaces.
5. Then I selected all blank spaces and put Null into them.
6. Then I highlighted all duplicates (using conditional formatting) of NHI because these should be unique.
7. Then I checked for any errors in the data (conditional formatting), there was none.
8. Then I set all names to proper string formatting.
9. Then substituted spaces between names into a comma only 1 instance so last names with spaces weren’t affected.
10. Then split the names into first and last for doctor patient and surgeon

## Convert excel data into csv file

* First, I had to delete the rows that had duplicates for the NHI (as this is the unique identifier for the patient), but I’ve held that data to enquire about it with the hospital.
* I created new sheets for each table in excel and copied all the data that is required. Including a primary key ID for all the tables except the patient (they have NHI), this took me awhile until I figured out how to do it.
* From here I wrote a little python script to convert each sheet into a csv file.

## Insert data into database

* Had to change date fields data type from smalldatetime to DATE to get the right output.
* I had to change the “Null” values in FSA column to be 1/1/1900 as it didn’t let me parse the data in. need to find out how to change back to null?

What went wrong but I fixed is I deleted the row of duplicate NHI numbers not realising it didn’t delete a cell that was a formula so I had to go back and do it again after and copy paste only values of the column.