UNSW Business School
School of Information Systems and Technology Management

# INFS2608 GROUP ASSIGNMENT COVER SHEET

Group (No/Name): T11A-02

Group Leader Name: Proteek Huq

Course: INFS2608 Database Management and Big Data Infrastructures

Tutorial time: Tuesday, 11am – 1pm

Tutor Name: T11A

Due Date and Time: 14 April 2023, 4pm (Moodle)

We declare that this assessment item is our own work, except where acknowledged, and has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

- Reproduce this assessment item and provide a copy to another member of the University; and/or,
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

We certify that we have read and understood the University Rules in respect of Student Academic Misconduct.

| zID | Name | Signed | Date |
|---|---|---|---|
| Z5416601 | Proteek Huq | P.H | 12/04/2023 |
| Z5421654 | Kathleen Ong | K.O | 12/04/2023 |
| Z5364294 | Hamza Khan | H.K | 12/04/2023 |
| Z5307203 | Ho, Sy Anh Quoc | Q.H | 12/04/2023 |

*The cover sheet is to be provided with any print or digital submission of the assignment. Failure to include a signed cover sheet will lead to a 5% penalty of the total marks available for the assignment. No marks will be released until a signed cover sheet is received.*

# Table of Contents

# 1 Introduction

WPH is a private hospital in the Eastern suburbs of Sydney, with a pristine reputation for delivering high quality medical services to its clients. The collection, application, and analysis of medical data is a fundamental aspect of the standards and efficiency of healthcare operations **(Victorian Quality Council Secretariat 2008)**. Tracking, monitoring, and Interpreting information exchanges is crucial for medical professionals to provide exceptional care **(DNSstuff 2021)**. Optimising the technical infrastructure that supports the collection and storage of data, would greatly improve the effectiveness and quality of operations as important information regarding staff, patients and procedures would be readily available and consistent. Furthermore, analysis of this data can provide managerial staff extremely valuable insights on hospital operations and help determine how resources should be allocated.

The purpose of this Proof-of-Concept Report is to demonstrate the feasibility for WPH to implement the Relational Model, Data Warehouse, and Business Intelligence models designed by Group T11A-02, to enhance the quality of its healthcare services.

Section 2 will outline improvements made to the Entity Relationship Diagram (ERD), justifying its ability to capture processes and data transactions that occur at WPH.

Section 3 will discuss the process of developing the Relational Schema of the Data warehouse, justifying the selection of tables and measures. The contextual suitability and effectiveness of this model will also be justified.

Finally, the potential for data management and big data analytics strategies will be explored and justified in Section 4 using Power BI data visualisations generated from a populated sample of the designed data warehouse. This section will demonstrate how Transaction data is created in the ERM, then updated in the data warehouse, where it can be retrieved using a materialised view that can then be used to generate data visualisations.

# 2 Entity-Relationship Diagram (ERD)

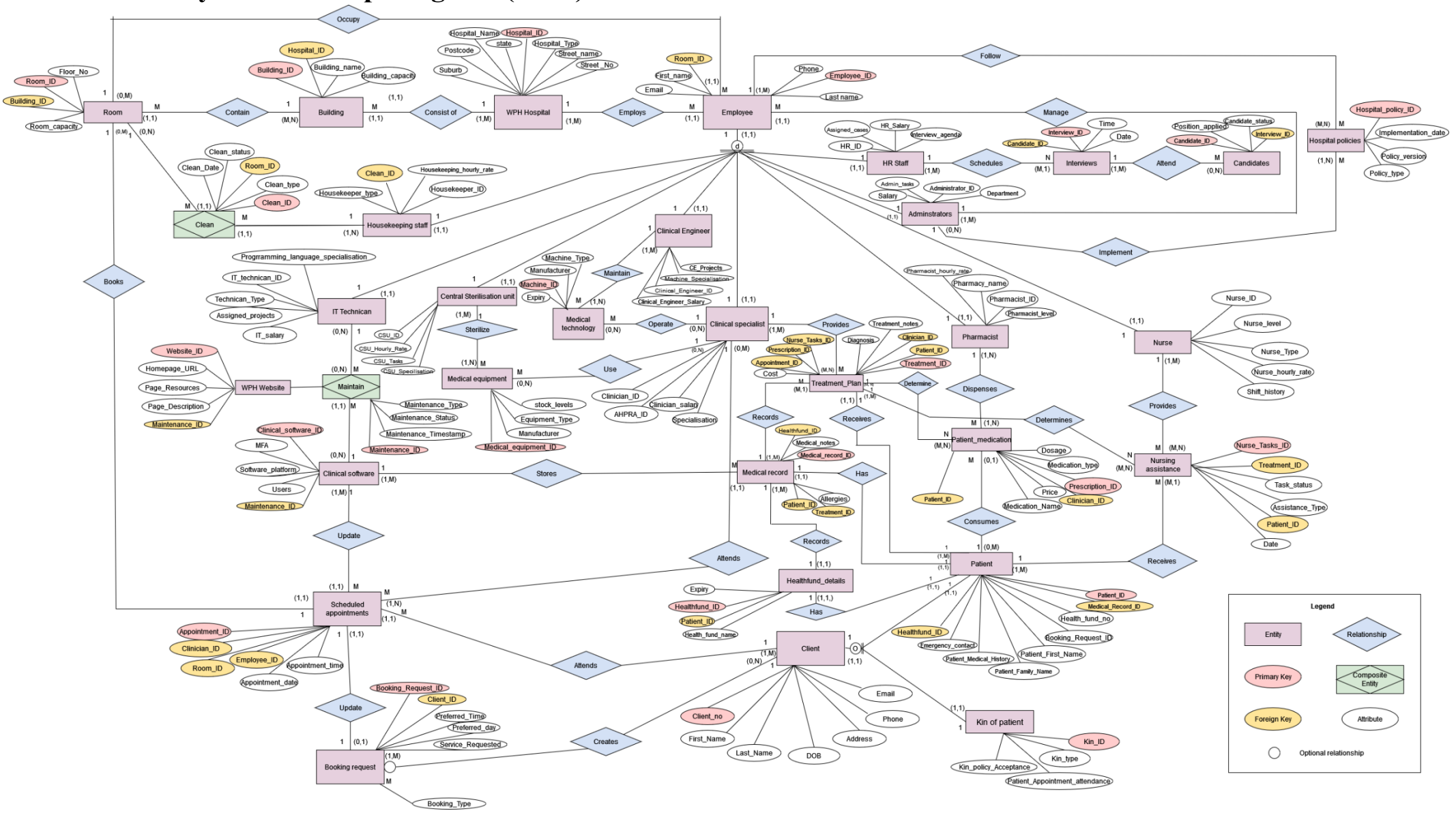





*Figure 2: Entity Relationship Diagram*

## 2.1 Changes Made to the ERD

An Entity Relationship Diagram (ERD) was reconstructed as shown in Figure 1 after reviewing feedback on the previous model. The purpose of the ERD was to convey the relationships between patients, staff and administrators within WPH. The diagram outlines the different types of employees and the actions that they are responsible for, e.g., pharmacists dispense medication. It also provides a comprehensive representation of the processes that occur when patients book appointments. E.g., A client creates a booking request to update their scheduled appointments, which are recorded on the clinical software that also records their medical records, including their treatment plans that the doctors assign them upon consultation. This treatment plan outlines the diagnosis, medication and nursing care that is to be assigned to the patient.

An issue regarding the previous diagram was its poor legibility. Group T11A-02 has taken several measures to rectify this problem. Firstly, the diagram was recreated on Draw.io, manipulating elements such as size, colour coding and a legend for improved clarity. It was also found that excessive entities were included in the previous ERD, which focused on portraying processes, rather than conveying how data is transferred and shared between entities.

For example, the previous ERD depicted several methods of creating booking requests (online booking, phone booking, front desk booking) as separate entities. This was later modified by connecting all the relevant attributes and relationships to a single “Booking Request” entity, whereby the booking method (phone, online, front desk) is identified by the attribute named “booking_type”.

Another issue of the previous ERD was the presence of incorrect foreign keys. This was rectified by ensuring that all entities were connected to the primary entities of their foreign keys. Furthermore, the ERD was revised to eliminate partial and transitive dependencies to maintain third normal form.

## 2.2 ERD Assumptions

When constructing our revised ERM, the following assumptions were adopted:

- Kin of patient can also be a patient.
- Patient_medication_type includes pills, ointments, etc.
- A patient can be assigned multiple treatment plans in one appointment.
- A clinical specialist can attend multiple appointments throughout the day; however they can only attend one at a time. Multiple clinical specialists may attend certain appointments such as surgeries.
- Employees such as clinical engineers, IT technicians, housekeeping staff, clinical specialists, and central sterilisation units (CSU) can have certain specialisations. For example:
    - A clinical engineer may specialise in maintaining a specific type of machinery such as X-rays or MRI machines.
    - One CSU staff may specialise in surgical equipment, whereas another member may be allocated to dental equipment.
    - Clinical Specialists work in certain fields such as Surgery, Oncology, Neurology, etc.
    - Housekeeping staff can include cleaners or landscapers.
    - IT technicians specialise in certain programming languages such as Python or JavaScript.
    - Maintenance_type includes software backups, updates, bug fixes, etc.

# 3 Relational Schema

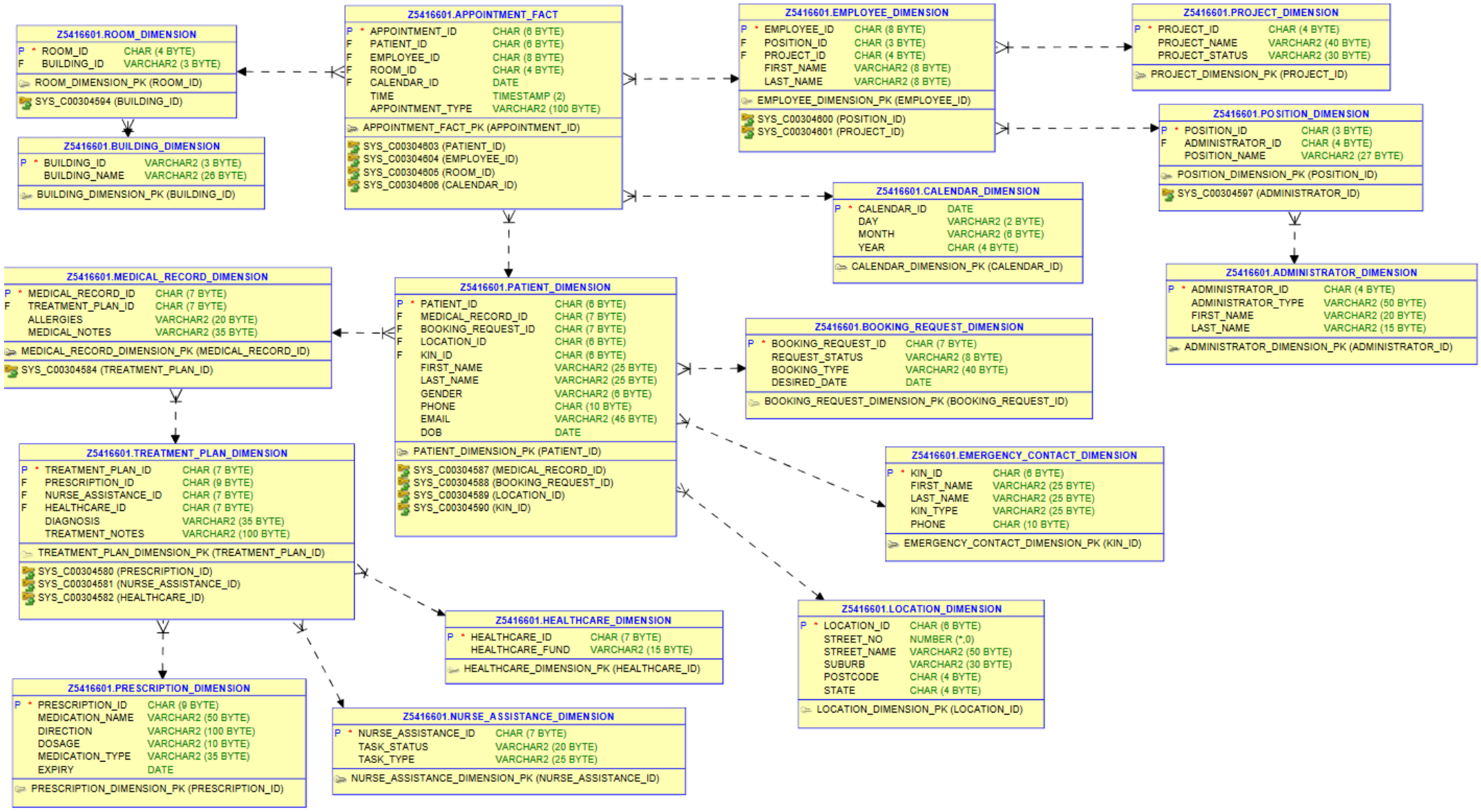



*Figure 3: Relational Schema*

### 3.1 Justifications for Relational Schema

Following the redevelopment of the ERD, Group T11A-02 modified its relational schema to reflect these changes and improve other aspects of the model so that data could be appropriately implemented in the data warehouse.

The intended purpose of the Schema was to convey the essential data needed to track patients and their personal data and medical information, as well as track employees and their certain roles and administrators.

Tables were developed in a manner such that they captured certain types of entities outlined in the ERD, they should be easily interpreted based on the names of tables and columns, have appropriate data types, and that tables facilitated functional links to other related dimensions. For example, while the ERD describes different types of employees such as “clinical specialists”, “pharmacists”, and “clinical engineers” as separate entities, the relational schema consolidates all employees within the “Employee_Dimension”, which is linked to other tables such as “Position_Dimension”, “Project_Dimension”, and “Administrator_Dimension” such that employees and their respective roles, projects and administrators are consistent and identifiable.

Feedback provided to team T11A-02, indicated the previous schema was too vague in terms of portraying specific data transactions, as the schema rather focused on depicting all the entities and processes included in the ERD. To produce a more data-centric model,

Appointment_Fact was selected as the fact table since it had the most useful variety of foreign keys such as Patient_ID, Room_ID, Calender_ID and Employee_ID, which connected the table to other dimensions that branched out into other useful information such as Building_ID, Medical_record_ID, Treatment_Plan_ID and Administrator_ID.

To support the branched nature of the different dimensions, a snowflake schema was the most appropriate type of layout to illustrate the connections between foreign keys **(R.F. Van Der Lans 2012)**. The successful implementation of this structure facilitated better data integrity by forcing constraints such as primary and foreign keys **(Watt 2014)**. After closely analysing the different types of relational schemas, it was concluded that the snowflake was the best schema to use because a galaxy schema proved too complex due to its nature of having two fact tables while the star schema did not allow for flexibility in regard to the analytical needs required from a normalised model **(What are the benefits and drawbacks of denormalizing data in star schema? 2023)**.

It was also found that the previous model included several dimensions such as website dimension, software dimension, medical machinery, and medical equipment dimension, that did not convey data that was particularly useful or relevant to healthcare processes. As a result, these dimensions were omitted from the schema so that the dimensions properly reflected the schema’s intended purpose.

The current snowflake schema also contains new tables such as the Calendar_Dimension. The calendar table allowed the team to track the dates of events occurring in the hospital such as appointments set by patients. These events can also be tracked or filtered using a hierarchy based on days, months, and years. A location dimension was also created to track the addresses of patients and implement a hierarchy of suburbs, postcodes, and states. A Project Dimension was also created to list different employee processes illustrated in the ERD. The project names would include actions such as providing GP consultations, hosting interviews, clinical software maintenance, cleaning, etc. Entities from the ERD such as Kin and patient medication were summarised in the Emergency_Contact and

Prescription dimensions respectively. These new names were applied to provide a more effective representation of data, and their respective transactions.

Furthermore, when creating tables, it was ensured all columns had appropriate measures and data types so that sample entries could be successfully imputed. To process this, a table was made to create sample entries for the various columns. A snapshot of this table is shown in Figure 3.1. For example, it was decided that Patient_ID would be measured as a unique 6-character chain of letters and numbers, therefore its data type was recorded as CHAR (6). Meanwhile, the name of different buildings could be more appropriately recorded using VARCHAR (26).

*Figure 3.1: Sample data entries*

| **Appointment_Fact** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Appointment_ID | A12873 | A36642 | A86956 | A93741 | A97884 | A31913 | A58085 | A92758 | A89005 | A89006 |
| Patient_ID | P12873 | P36642 | P86956 | P93741 | P97884 | P31913 | P58085 | P92758 | P89005 | P65209 |
| Employee_ID | EMP1001 | EMP1003 | EMP1004 | EMP1005 | EMP1006 | EMP1007 | EMP1008 | EMP1009 | EMP1010 | EMP1001 |
| Room_ID | R167 | R158 | R344 | R421 | R621 | R324 | R188 | R666 | R100 | R972 |
| Calender_ID | 12/03/2023 | 09/08/2022 | 20/06/2023 | 27/04/23 | 01/05/2023 | 09/04/2023 | 13/05/2023 | 18/04/2023 | 03/06/2023 | 06/06/2023 |
| Time | 11:00 | 10:00 | 14:00 | 14:00 | 9:00 | 18:00 | 15:00 | 12:30 | 17:00 | 15:00 |
| Appointment_Type | GP consultation | Dermatologist Appointment | Surgery | Mammogram | Cancer Screening | Emergency Care | Neurologist Appointment | Endocrinologist appointment | Monthly check up | Monthly Check up |
| | | | | | | | | | | |
| **Calender_Dimension** | | | | | | | | | | |
| Calender_ID | 12/03/2023 | 09/08/2022 | 20/06/2023 | 27/04/23 | 01/05/2023 | 09/04/2023 | 13/05/2023 | 18/04/2023 | 26/06/2023 | 06/06/2023 |
| Day | 12 | 9 | 20 | 27 | 1 | 9 | 13 | 18 | 26 | 6 |
| Month | March | August | June | April | May | April | May | April | June | June |
| Year | 2023 | 2022 | 2023 | 2023 | 2023 | 2023 | 2023 | 2023 | 2023 | 2023 |
| | | | | | | | | | | |
| **Room_Dimension** | | | | | | | | | | |
| Room_ID | R167 | R158 | R344 | R421 | R621 | R324 | R188 | R666 | R100 | R972 |
| Building_ID | B1 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B2 |
| | | | | | | | | | | |
| **Building_Dimension** | | | | | | | | | | |
| Building_ID | B1 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B2 |
| Building_Name | The Blue Centre | Birkley Building | WPH Main Building | WPH centre of Radiology | WPH Cancer Research Clinic | WPH Emergency Clinic | Goodwill Centre | Green Centre | WHP North Wing | Chleo Cox Centre |

# 4 Business Intelligence (PowerBI)

WPH's big data strategy to improve the quality of its services will likely involve the extraction of insights into patient's personal details and medical circumstances, as well as the development of data visualisations that can assist with streamlining processes or identifying specific information.

Once the Data warehouse was populated, a materialised view was created as shown below in Figure 3, the code for the materialised view is also included in the appendix. This materialised view was then exported and implemented in PowerBI, where it was used to generate the data visualisations below.

Analysis of big data using tools such as Power BI can provide managerial staff valuable insights that enable them to make informed decisions that can optimise operational efficiency and **quality (T. Choi, S. W. Wallace, Y. Wang, 2017)**.

*Figure 4: Materialised view output*

| | PATIENT_ID | GENDER | HEALTHCARE_FUND | LOCATION_ID | SUBURB | POSTCODE | STATE | ALLERGIES | DIAGNOSIS | MEDICATION_NAME | MEDICATION_TYPE | APPOINTMENT_TYPE | CALENDAR_ID | EMPLOYEE_ID | POSITION_NAME | ADMINISTRATOR_ID | ADMINISTRATOR_TYPE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | P12873 | Male | Medibank | L12873 | Kogarah | 2217 | NSW | Nut | Influenza | Zanamivir | Tablet | GP Consultation | 12/MAR/23 | EMP1001 | General Practitioner | AD10 | GP Administrator |
| 2 | P31913 | Male | Bupa | L31913 | Bonang | 3888 | VIC | N/A | Pneumonia | Amoxicillin | Capsule | Emergency Care | 09/APR/23 | EMP1007 | Pulmonologist | AD15 | Pulmonology Administrator |
| 3 | P36642 | Male | Bupa | L36642 | Coogee | 2034 | NSW | Shellfish | Eczema | Triamcinolone | Topical Ointment | Dermatologist Appointment | 09/AUG/22 | EMP1003 | Dermatologist | AD11 | Dermatology Administrator |
| 4 | P58085 | Female | HCF | L58085 | Lindfield | 2070 | NSW | Gluten | Parkinson Disease | Levodopa | Tablet | Neurologist Appointment | 13/MAY/23 | EMP1008 | Neurologist | AD16 | Neurology Administrator |
| 5 | P65209 | Male | HCF | L65209 | Abbotsbury | 2176 | NSW | Dairy | Hypertension | Telmisartan | Tablet | Monthly Check up | 06/JUN/23 | EMP1001 | General Practitioner | AD10 | GP Administrator |
| 6 | P86956 | Male | Nib | L86956 | Randwick | 2031 | NSW | N/A | Appendicitis | Lidocaine | Injection | Surgery | 20/JUN/23 | EMP1004 | Surgeon | AD12 | Surgery Administrator |
| 7 | P89005 | Male | Bupa | L89005 | Ingleburn | 2565 | NSW | N/A | Gastric Ulcer | Pantoprazole | Tablet | Monthly check up | 26/JUN/23 | EMP1010 | Gastroenterologist | AD18 | Gastroenterology Administrat |
| 8 | P92758 | Male | Medibank | L92758 | Westmead | 2145 | NSW | N/A | Thyroid disease | Levothyroxine | Tablet | Endocrinologist appointment | 18/APR/23 | EMP1009 | Endocrinologist | AD17 | Endocrinology Administrator |
| 9 | P93741 | Male | HCF | L93741 | Matraville | 2036 | NSW | N/A | Breast Cancer | Chemotherapy | Chemotherapy | Mammogram | 27/APR/23 | EMP1005 | Radiologist | AD13 | Radiology Administrator |
| 10 | P97884 | Female | HCF | L97884 | Braddon | 2612 | ACT | Latex | Lymphoma | Chemotherapy | Chemotherapy | Cancer Screening | 01/MAY/23 | EMP1006 | Oncologist | AD14 | Oncology Administrator |

*Figure 4.1: Map highlighting the different suburbs that patients live in*

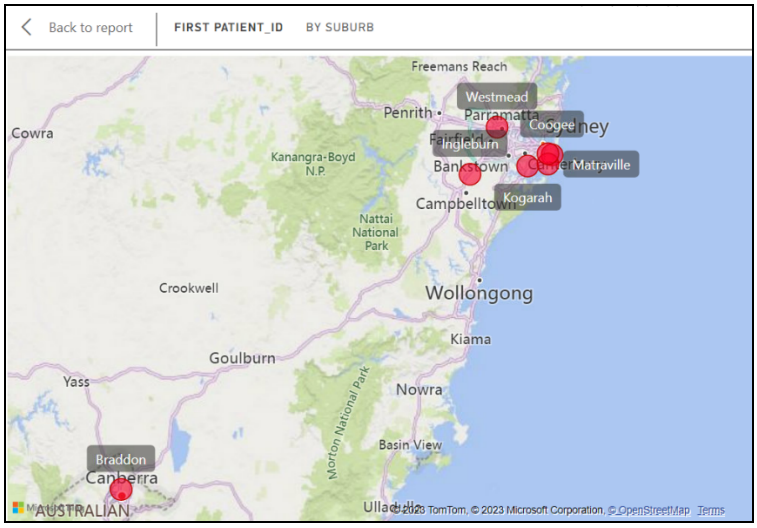



*Figure 4.2- Table of number of patients living in each state*

| NUMBER OF PATIENTS | STATE |
|---|---|
| 1 | ACT |
| 8 | NSW |
| 1 | VIC |
| **10** | |

WPH serves patients from a range of suburbs and states across Australia. This can be seen in figure 4.1 whereby most patients live within inner Sydney, meanwhile, a patient lives in the suburb of Braddon, Canberra. Additionally, the distribution of patients per state is indicated in Figure 4.2, which uses the measure: 'Number of Patients = COUNT('INFS2608 MV'[PATIENT_ID])'. Analysing the spread of patient's addresses, can enable a better understanding of patient demographics.

*Figure 4.3- Pie Chart of patients and gender*

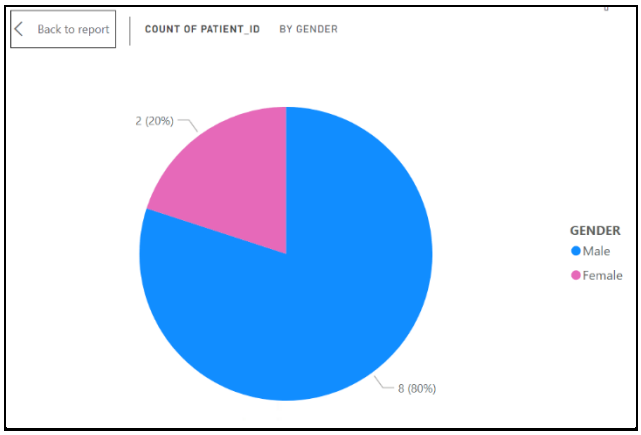



Figure 4.3 demonstrates the high male to female gender ratio of patients at WPH. This data addresses potential inequalities in the healthcare industry. Managerial staff may use this data as a basis to investigate the reason behind low levels of female patients, or to adjust its resources in accordance with these proportions.

*Figure 4.4- Table of Patients' medical information*

| PATIENT_ID | ALLERGIES | DIAGNOSIS | MEDICATION_NAME |
|---|---|---|---|
| P12873 | Nut | Influenza | Zanamivir |
| P31913 | N/A | Pneumonia | Amoxicillin |
| P36642 | Shellfish | Eczema | Triamcinolone |
| P58085 | Gluten | Parkinson Disease | Levodopa |
| P65209 | Dairy | Hypertension | Telmisartan |
| P86956 | N/A | Appendicitis | Lidocaine |
| P89005 | N/A | Gastric Ulcer | Pantoprazole |
| P92758 | N/A | Thyroid disease | Levothyroxine |
| P93741 | N/A | Breast Cancer | Chemotherapy |
| P97884 | Latex | Lymphoma | Chemotherapy |

Figure 4.4 provides a consolidated view of each patient and their corresponding allergies, diagnosis, and current medication. This graph is useful for gaining a holistic insight of each patient's medical situation.

*Figure 4.5 Total appointments each month*

| NUMBER OF APPOINTMENTS | Year | Month |
|---|---|---|
| 1 | 2022 | August |
| 1 | 2023 | March |
| 3 | 2023 | April |
| 2 | 2023 | May |
| 3 | 2023 | June |
| **10** | | |

Figure 4.5 calculates the number of appointments for each month of the year. April 2023 and June 2023 are the busiest months while August 2022 as well as March 2023 are the least busy. The data may help WPH decide which time of year they should increase their resources to manage patient inflows.

*Figure 4.6 Total patients per Healthcare Fund*

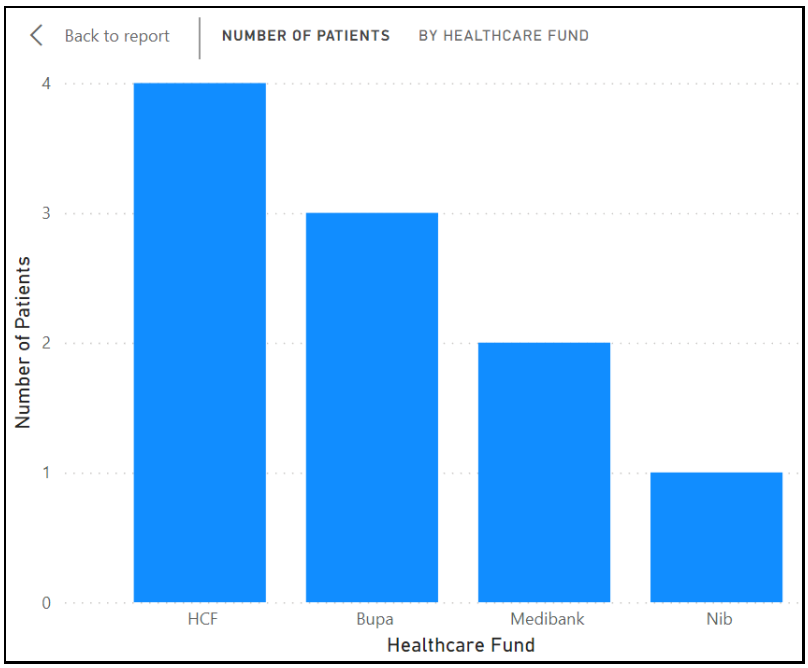
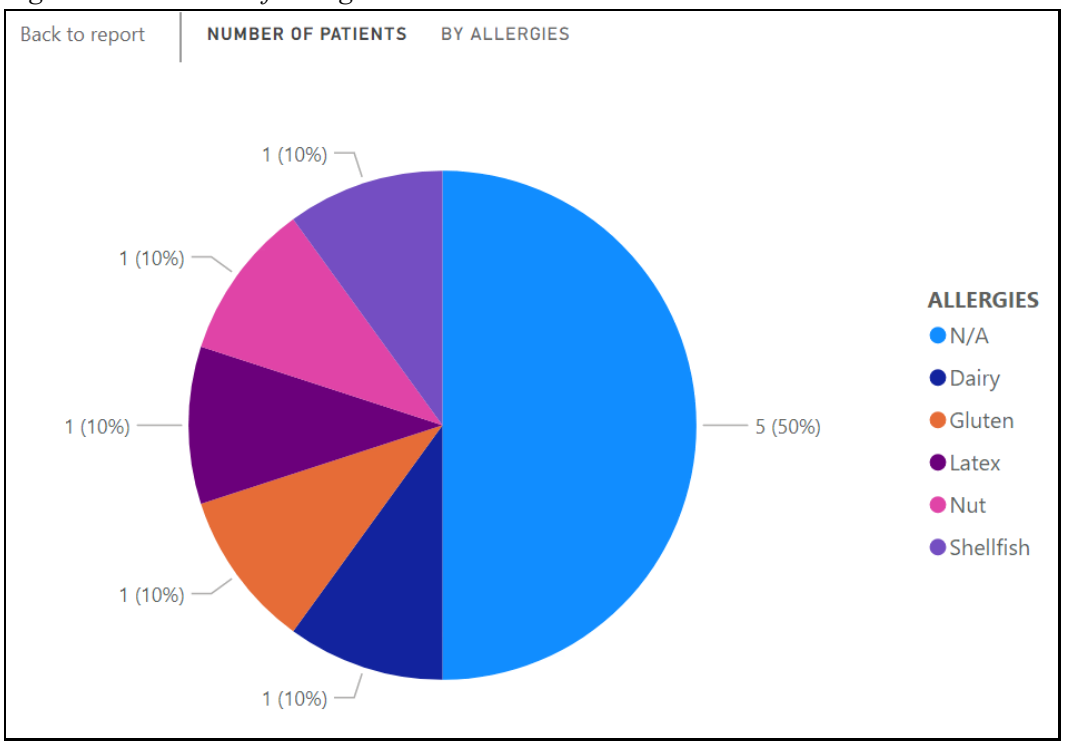



Figure 4.6 shows the number of patients for each healthcare fund. The most popular healthcare fund in HCF meanwhile Nib is least popular. WPH may use this data to negotiate favourable reimbursements rates for medical services.

*Figure 4.7 Pie chart of Allergies*




The pie chart showcases the different allergies present among patients of WPH. This information can be used to determine what precautions need to be taken at WPH to cater for such patients.

# 5 Conclusion

In conclusion, this report clearly demonstrates that the database designed by Group T11A-02 is a very appropriate model for WPH to implement. The Entity Relationship Diagram conveys a comprehensive, data-centric understanding of the relationships and processes existing in WPH. This is reflected in the Relational Schema which effectively connects the different dimensions of the database. The functionality of this database was validated through the computation of sample entries and a Materialized view. The creation of various data visualisations proved that WPH management can leverage the database to develop big data strategies to optimise operations.

# 6 Appendix

**Materialised view:**

```
CREATE MATERIALIZED VIEW WPH_APPT_MV

BUILD IMMEDIATE

AS SELECT

a.PATIENT_ID,

p.GENDER,

h.HEALTHCARE_FUND,

p.LOCATION_ID,

l.SUBURB,

l.POSTCODE,

l.STATE,

mr.ALLERGIES,

tp.DIAGNOSIS,

pres.MEDICATION_NAME,

pres.MEDICATION_TYPE,

a.APPOINTMENT_TYPE,

c.CALENDAR_ID,

e.EMPLOYEE_ID,

j.POSITION_NAME,

ad.ADMINISTRATOR_ID,

ad.ADMINISTRATOR_TYPE

FROM APPOINTMENT_FACT a

JOIN PATIENT_DIMENSION p

ON (a.PATIENT_ID = p.PATIENT_ID)

JOIN EMPLOYEE_DIMENSION e

ON (a.EMPLOYEE_ID = e.EMPLOYEE_ID)

JOIN CALENDAR_DIMENSION c

ON (a.CALENDAR_ID = c.CALENDAR_ID)
```

```
JOIN LOCATION_DIMENSION l

ON (p.LOCATION_ID = l.LOCATION_ID)

JOIN MEDICAL_RECORD_DIMENSION mr

ON (p.MEDICAL_RECORD_ID = mr.MEDICAL_RECORD_ID)

JOIN TREATMENT_PLAN_DIMENSION tp

ON (mr.TREATMENT_PLAN_ID = tp.TREATMENT_PLAN_ID)

JOIN HEALTHCARE_DIMENSION h

ON (tp.HEALTHCARE_ID = h.HEALTHCARE_ID)

JOIN PRESCRIPTION_DIMENSION pres

ON (tp.PRESCRIPTION_ID = pres.PRESCRIPTION_ID)

JOIN POSITION_DIMENSION j

ON (e.POSITION_ID = j.POSITION_ID)

JOIN ADMINISTRATOR_DIMENSION ad

ON (j.ADMINISTRATOR_ID = ad.ADMINISTRATOR_ID);
```

# 7 References


DNSstuff 2021, How Important Are Databases in the Healthcare System? - DNSstuff, Software Reviews, Opinions, and Tips - DNSstuff, viewed 9 April 2023, <https://www.dnsstuff.com/how-important-database-healthcare>.

R.F. Van Der Lans 2012, Snowflake Schema - an overview | ScienceDirect Topics, www.sciencedirect.com, viewed 8 April 2023, <https://www.sciencedirect.com/topics/computer-science/snowflake-schema>.

Victorian Quality council Secretariat 2008, A guide to using data for health care quality improvement, viewed 11 April 2023, <https://aci.health.nsw.gov.au/__data/assets/pdf_file/0006/273336/vqc-guide-to-using-data.pdf>.

Watt, A 2014, Chapter 9 Integrity Rules and Constraints, Opentextbc.ca, BCcampus, viewed 14 April 2023, <https://opentextbc.ca/dbdesign01/chapter/chapter-9-integrity-rules-and-constraints/>.

What are the benefits and drawbacks of denormalizing data in star schema? 2023, Linkedin.com, viewed 14 April 2023, <https://www.linkedin.com/advice/0/what-benefits-drawbacks-denormalizing-data-star-schema-skills-olap>.

T. Choi, S. W. Wallace, Y. Wang, 2017, ‘Big Data Analytics in Operations Management’, Volume27, Issue10, ‘Special Issue on Big Data in Supply Chain Management’, https://doi.org/10.1111/poms.12838