AHRI health informatics internship – assessment exercises

Question 1

A study was conducted in AHRI's demographic surveillance area in rural KwaZulu Natal in 2018-2019. All residents aged 16 years and over were invited to attend a general health screen for hypertension at a mobile clinic. Those who were diagnosed with hypertension were referred to care at one of the primary health care facilities in the area.

Researchers were interested in whether participants who were referred to care attended a clinic after referral. They were also interested in the clinic attendance patters among people who participated in the study. The clinics in the area collect routine data on the reason for visit, for all individuals who attend the clinic.

We have given you four datasets (described in the table below), containing information on a subset of participants in the study. We would like you to do the following:

1. Check the datasets for errors and list any issues that you find.
2. Visualise the data – this can be with graphs, charts, etc. – whatever you feel is best
3. Analyse the data to answer the following questions:
    a. How many of the participants who were diagnosed with hypertension attended a clinic?
    b. How many had hypertension as the reason for their visit?
    c. How many other participants attended a clinic?
    d. What were the most common reasons for visits?

As part of the interview, we would like you to give a short presentation (10 minutes maximum) of your findings.

Name	Description
Participants.csv	Identifying information for subset of participants in the health screen study: • id_new: participant ID • sex • date of birth
Health_screen.csv	Data gathered during the health screening study • id_new: participant ID • date_screen: date of screening visit • systolicBP: systolic blood pressure • diastolicBP: diastolic blood pressure • bmi: body mass index • smokecat: smoking status • BPdiag: diagnosed with hypertension at screening visit (yes/no)
Clinic_visits.csv	Clinic visit information for subset of participants in health screen study • id_new: participant ID • visitdate: date of clinic visit • visitreason: Reason for clinic visit
Clinic_codes.csv	Clinic ID code for all visits in 'clinic_visits' dataset

Question 2

Researchers collected data from 2,500 hospitals to find out whether they were meeting their target of seeing at least 93% of cancer referral patients "on time". They extracted data for two years: 2020 and 2021. The graph below shows their results. They concluded that:

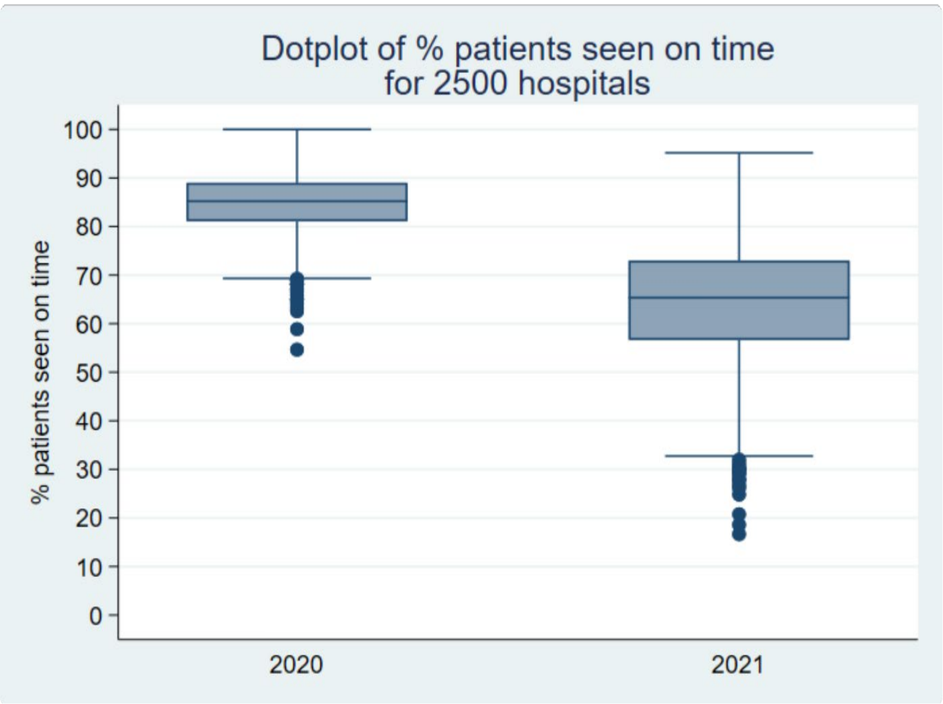
- No hospital managed to see all their cancer referral patients on time in 2021.
- In 2021, some hospitals managed to see fewer than 20 patients on time.
- On average, the percentage of patients seen on time was lower in 2021 than 2020.
- The backlog caused by the COVID-19 pandemic has reduced hospitals' ability to handle their patient load.

a. For each of the four researcher statements, state whether it is supported, partially supported, or not supported by the graph. Please give a short explanation to justify your response (maximum 2 sentences for each point).

b. Estimate the median “% seen on time” for each year and the approximate difference between years.

c. Discuss one alternative explanation for the observed change other than the COVID-19 backlog.

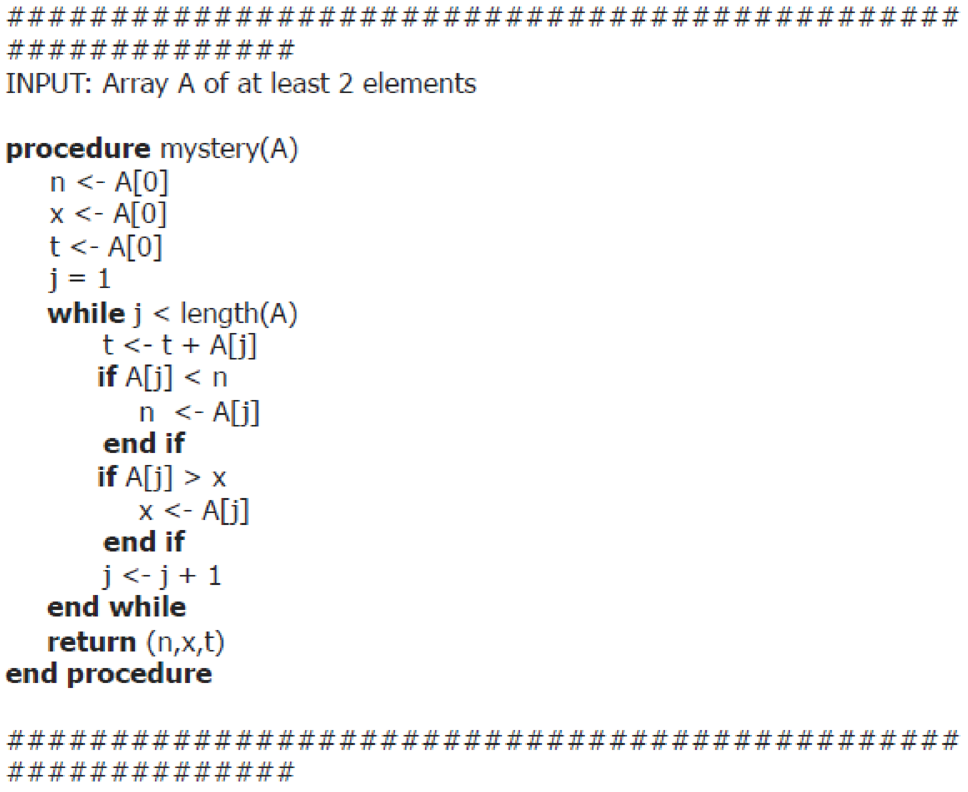
d. Suggest one additional analysis or dataset that would strengthen or challenge the researchers’ conclusion.



Question 3.

In health informatics, we often need to process and summarise large sets of clinical or patient-generated data. The pseudocode below describes an algorithm called mystery(A). It takes an array A as input — for example, a list of daily measurement values collected from a wearable device.

Note that the array A is indexed from 0 rather than 1 (as you might be more used to in a mathematical setting), i.e. A = [A[0], A[1],...] rather than A = [A[1], A[2],...].



```
##############################################
##############
INPUT: Array A of at least 2 elements

procedure mystery(A)
    n <- A[0]
    x <- A[0]
    t <- A[0]
    j = 1
    while j < length(A)
        t <- t + A[j]
        if A[j] < n
            n  <- A[j]
        end if
        if A[j] > x
            x <- A[j]
        end if
        j <- j + 1
    end while
    return (n,x,t)
end procedure

##############################################
##############
```

Imagine that A contains daily resting heart rate (RHR) readings (in beats per minute) for one patient over several weeks. You've been asked to interpret and potentially adapt this algorithm for use in a patient monitoring dashboard.

a. In plain language, explain what the mystery algorithm calculates.

b. What do each of the three returned values (n, x, t) represent in this patient-monitoring context?

c. Demonstrate the algorithm's behaviour on the following dataset:

   A = [72, 74, 70, 69, 76]

   Show the intermediate updates to n, x, and t after each loop iteration.

d. Modify the pseudocode (or write real code in a language you prefer) so that it also computes:

   - The average heart rate across the period, and
   - A flag if any reading is more than 15% above the average