Solution

Exercise 1

The main points to consider are the following:

Randomized trial: The best solution to a question about effect of an intervention, if it can be carried out. Due to the randomization, we avoid confounding problems.

Cohort study: A good alternative in situations where a proper randomized study cannot be carried out, but as it is an observational study, confounding problems may occur.

The following is more directly related to the current situation and can be seen as additional information:

The research question here is about the effect of physical activity on the risk of hypertension. This indicates that we would have to recruit people without hypertension and follow them over time. For a randomized trial, this would mean to recruit a high number of healthy individuals and randomize them to some exercise program or not, which they would have to follow over a long period of time (years), and then count how many subjects develop hypertension. This is probably unfeasible in practice. An alternative, which doesn't directly answer the research question, is to include hypertensive patients, randomize them to the same type of intervention, follow them over time (not necessarily so long here) and observe whether blood pressure is reduced.

In a cohort study, the recruitment would be as above; include healthy individuals, register their physical activity habits and follow them over time. In addition, potential confounding factors would need to be registered, typically lifestyle factors. Note that it may not be obvious how to register physical activity.

Exercise 2

A number of solutions are relevant, as you have a high number of variables to choose among. We are primarily interested in how you present the variables you choose, not so much which variables you choose. Categorical variables (e.g. gender, marital status, work status …) should be presented as frequencies and percentages, while numerical variables should be presented by mean / median with some measure of variation (standard deviation, min-max, …). Regarding the choice between mean and median that is mentioned in the exercise, one might argue that a numerical variable with a highly skewed variable should be presented by a median value rather than a mean.

When it comes to choice of variables, notice that the exercise asks you not to present results related to the purpose of the study. Thus, it would make sense to avoid describing the sample through variables that are part of the outcome (e.g. pain measurements). However, this will not be emphasized in the evaluation.

Exercise 3

We have focused quite a bit on the difference between categorical data and numerical data, and we have said that it is important to be able to distinguish between the two types of data, as this also dictates the choice of statistical methods. This exercise is about understanding whether the pairing of data and method makes sense. The variable that is analyzed (How is your musculoskeletal pain ..) is clearly a categorical variable, while the method (t-test) is a method that is meant for numerical data. Thus, we would not trust this conclusion.

Exercise 4

a) The estimated risk of preeclampsia is given by 141/1500 = 0.094 (or 9.4%).
A 95% confidence interval is given by

$$0.094 \pm 1.96\sqrt{\frac{0.094(1-0.094)}{1500}},$$

which gives an interval (0.079, 0.109).

b) $\widehat{RR} = \frac{29/200}{112/1300} \approx 1.7$. This suggests that the risk of preeclampsia is increased by 70% among mothers ≥ 35 years compared to those below 35. Another way of saying this is that the risk is 1.7 times as high among those ≥ 35 years compared to those below 35.

c) There is a significant difference between the two groups, as the interval does not include the null-value, which is '1' for a RR.

Exercise 5

We have an estimated effect of treatment = -1.84 (the regression coefficient for Trt) with a corresponding 95% confidence interval (-4.27, 0.59). The effect is non-significant by a p-value = 0.14 (0.136). This is the essential information to answer the question in the exercise.

The interpretation of this is that, for given baseline value (i.e. if they all have the same blood pressure at baseline), the estimated effect of one of the treatments (compared to the other) is to reduce diastolic blood pressure by on average 1.84 mmol/l. This is a very small improvement, and it makes sense to conclude with 'no significant effect'.