**Question 2 (4 points)**

1. Write a brief description of the study by Jabbour in your own words, including the following:

(a) Research objectives

(b) Patient population recruited for the study

(c) Study design

(d) Primary outcome

(e) Result for the primary outcome

The study conducted by Jabbour and colleagues, published in JAMA 2024, was a phase 3 global randomized trial in newly diagnosed Ph+ ALL patients with two frontline methods: ponatinib and imatinib. The objective was to compare pontinib vesus imatinib, both given with reduced-intensity chemotherapy. 245 patients from 77 sites were randomized in a 2:1 ratio to receive ponatinib or imatinib therapy. The primary outcome was MRD-negative complete remission at the end of induction. It was defined as morphological remission sustained for at least 4 weeks plus central laboratory MRD negativity by RT-qPCR. Results showed ponatinib achieved significantly better outcomes: 34.4% of patients reached MRD-negative CR compared with 16.7% in the imatinib group, with a risk difference of 18% and a P value of 0.002. In summary, ponatinib combined with low-intensity chemotherapy is a better frontline option than imatinib for Ph+ ALL.

1. Which variables are continuous?

Age

1. Which variables are unordered categorical variables?

Sex, Central nervous system disease/extrameduallary disease, BCR::ABL1 dominant isoform(p190, p210, atypical, undermined or not tested), cardiovascular comoribidities Hypertension (Yes/No)， Diabetes (Yes/No), Obesity (Yes/No), Dyslipidemia (Yes/No), History of smoking (Yes/No)

1. Which variables are ordered categorical variables?

ECOG performance status (0, 1, 2) , age group stratification (others, ≥60 years), Number of cardiovascular comorbidities (≥1, ≥2)

1. Why does the table not have an “overall” column that shows descriptive statistics for all 245 patients?

That’s because the purpose of Table 1 in an RCT is to show comparability between randomized groups. The randomization already ensures baseline balance on average; an “overall” column adds little value and could be misleading. Moreover, according to CONSORT guidance, baseline data should be reported by group only, not overall.

1. Why are there no p-values in Table 1?

Because randomization ensures balance, hypothesis testing at baseline is inappropriate.

**Question 3 (4 points)**

1. Review the CONSORT diagram in Figure 1 from the article by Jabbour and write a brief summary of what you see. Describe reasons why patients were excluded (see notes under Table 1 for additional details) and what, if anything, might cause you to be concerned about generalizing results from those who were analyzed to those who were enrolled and randomized. After doing this, consider what the target population for the study is and list any concerns you have about generalizing the study results to that population.

The trial randomized 245 patients (164 ponatinib, 81 imatinib). One ponatinib patient died of COVID-19 before treatment. Ten ponatinib and three imatinib patients were excluded from the primary analysis due to atypical or untested BCR::ABL1 transcripts. This left 154 ponatinib and 78 imatinib patients for the main endpoint. These exclusions may limit generalizability: if excluded patients differ systematically, results may not apply to them. The intended target population is all adults with newly diagnosed Ph+ ALL, but the analyzed population excludes those with rare transcript variants.

1. Summarize Altman’s argument about not performing statistical testing on Table 1 in a randomized trial. *Note: Your instructors are aware that understanding Altman’s argument on a deep level assumes familiarity with material you haven’t yet covered in BIOSTAT 701. Don’t worry about that for now. Simply do your best to describe Altman’s argument in a non-technical fashion using what you currently know about statistics.*

Altman argued that p-values are not meaningful for baseline comparisons in randomized trials. Randomization guarantees comparability on average, so any observed imbalance is just chance. A small p-value does not mean randomization failed, and a large p-value does not prove groups are identical. Therefore, Table 1 should report descriptive summaries only.

1. Shift your attention back to the ultra-running study for a moment. Recall that this is an observational study instead of a randomized trial, and that our objective (which we are slowly building towards) is to run a simple linear regression predicting best running time based on emotional intelligence. In the last exercise you created a table that compared the 73 participants who would be excluded from the simple linear regression with the 211 participants who would be included. The table had a very similar structure to the typical Table 1 from a randomized trial, i.e., that it compares baseline characteristics between two independent groups. The purpose of the table was different, however. In other words, the exercise with the ultra-running study was intended to identify selection bias, i.e., systematic differences between the analysis set and the entire sample. This is contrasted with the purpose of Table 1 in a randomized trial, which is to assess imbalance in baseline factors that might result in another kind of bias called confounding, which as Altman demonstrates in his paper, can happen even in a randomized trial. This phenomenon occurs specifically when the imbalance is on factor(s) that are also associated with the outcome.

You already answered a question above about why it doesn’t make sense to use p-values to asses baseline differences between randomized groups in a clinical trial. Explain why it also doesn’t make sense to use p-values in the ultra-running study context, where our objective was to assess selection bias.

In the ultra-running study, the baseline table compared participants included vs. excluded from regression. The goal was to assess selection bias. Here, p-values are not useful either: large samples can make trivial differences significant, while small samples may miss important imbalances. What matters is the size and direction of differences, not arbitrary significance.

**Question 4 (4 points)**

A table of numbers and letters

AI-generated content may be incorrect.

**Question 5 (4 points)**

Answer the following questions about the Table 1 that you generated. 3

1. Would you reject the null hypothesis of no difference between the treatment arms for any of the baseline characteristics in Table 1 (using a 5% alpha level)?

No. All baseline p-values are ≥0.05; thus I would not reject the null for any variable.

1. If you were to interpret the p-value as continuous, rather than using a threshold value like 5%, then are there any factors for which the p-value is suggesting the evidence leans in favor of the alternative hypothesis (i.e., that the distribution of a factor is actually different in patients assigned to Ponatinib vs. Imatinib)? Discuss why this doesn’t make any sense in this context. Refer to Altman’s paper as a guide as you think about this.

The smallest p-values are for Diabetes (p=0.071), Gender (p=0.094), and ECOG (p=0.115), so these would “lean” most toward imbalance. But per Altman, baseline testing in RCTs is misleading: randomization already makes any differences chance findings; small p’s don’t mean randomization failed, and large p’s don’t prove equality.

1. Are there any factors that are imbalanced between the groups when you look at the standardized mean differences (use 0.2 as a threshold)? List the factors and describe the imbalance that you see.

Yes: Gender SMD=0.247, ECOG SMD=0.280, Diabetes SMD=0.247 exceed 0.20. Direction: more females and higher ECOG in the ponatinib arm; diabetes more common in ponatinib (means ≈0.58 vs 0.46).

1. Why do you think the standardize mean differences are a more satisfactory approach to assessing imbalance than using p-values?

SMDs quantify magnitude on a common scale and are insensitive to sample size. p-values mix effect size with n (big n → trivial differences “significant”; small n → important differences non-significant). SMDs allow consistent thresholds (0.1/0.2/0.5) across variable types.

1. Suppose that the factors you identified as being imbalanced were also strongly associated with the primary outcome of the study. Using Altman’s argument, what might you consider doing if you were the statistician for the study?

Treat them as potential confounders: prespecify or perform covariate-adjusted analyses (e.g., regression/CMH with covariates), consider stratification, and run sensitivity analyses. Report adjusted effect estimates with CIs alongside the primary analysis.

1. Suppose that the factors you identified as being imbalanced had no association with the primary outcome of the study. Again, referring to Altman’s paper, is the imbalance a concern and what, if anything, would you consider doing about it?

Then they are not confounders; imbalance is usually not a validity threat. You may optionally adjust for precision, but no adjustment is required for bias control.