1. **Q1. Is there a statistically significant difference in the expression levels of the 4 immune cell markers pre-treatment and post-treatment (analyze each marker separately)?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **IL1B** | **CX3CL1** | **TNFA** | **CCL20** |
| **Distribution** | Non normal | Non normal | Non normal | Non normal |
| **Test used** | Wilcox | Wilcox | Wilcox | Wilcox |
| **Statistically different?** | Yes | Yes | Yes | Yes |

* First the data was visualised to aid in assessment of distribution.
* Text

  Description automatically generated
* The variables were subbed in for pre and post treatment.

|  |  |
| --- | --- |
| Pre-Treatment Arm: | Post-Treatment Arm: |
|  |  |
|  |  |
| * On visual inspection the histogram bars do not completely follow the fitted line of distribution. We can see there is a drop in values between 4 and 4.5 where a peak or equal level should really lie. * On the QQ plot the points match the line of normal distribution quite well with few deviating. The hypothesis test confirms the data is not normally distributed as the Shapiro test gave us a non-significant p value. | * The post treatment data is not normally distributed and is considered right skewed. The histogram bars do however follow the fitted line of distribution rather well. We can also see how the tail is skewed on the QQ plot as the points drift away from the line of normal distribution(red). A significant hypothesis test also confirms this. * Since some of the data is nonparametric the Wilcox test will be used to compare the medians to identify a significant difference. |
| Shapiro p value: p<0.0564 | Shapiro p value: p<5.352e-11 |
| * A box plot can also aid us visually.      * We can see visually on both the histograms above and the box plots that there is an apparent difference in the expression level of the marker pre and post treatment in both the medians and the range. * The pre data trends closer together while the post data is lower overall but contains outliers. * We will confirm this with a hypothesis test. | |
| * Text    Description automatically generated * We let paired=TRUE so R knows the samples are matched. * This gives us a significant p value(<0.05) so we can say there is a statistically significant difference between the medians of IL1B pre and post treatment. | |

* The initial steps to identify normality were repeated for CX3CL1, the same code was used and the variables subbed in.

|  |  |
| --- | --- |
| Pre-Treatment Arm: | Post-Treatment Arm: |
|  |  |
|  |  |
| * In general, the bars on the histogram follow the line with a peak and two tails, however the peak is above the expected probability of finding an expression level of this value. * Very few points lie on the line of normal distribution in the QQ plot, we can see the peak of the data as the points dip below towards the centre of the line. * However, the hypothesis test returned a significant p value(<0.05) indicating that this data is not normally distributed. | * Post-Treatment: * It is clear from the histogram that this data is not normally distributed. It appears to be multimodal with a peak in the lower and higher ends of the data. * The QQ plot also shows deviations from the line of normal distribution in the upper range as seen in the histogram. This second peak could be a treatment trend or random outliers. * The hypothesis test of significance returned a significant p value(<0.05) clearly indicating the data isn’t normally distributed. |
| Shapiro p value: p<0.03667 | Shapiro p value: p<4.747e-09 |
| * There is a slight difference in the medians of the data on visual inspection. However the post data median could be higher because of the number of outliers in the post treatment data influencing its position. * Since the data is nonparametric, I ran the Wilcox test again to compare the medians of the paired samples. | |
| * A picture containing text    Description automatically generated * The p value return was significant(<0.05) thus we can say there is a significant difference between the medians of the ranked data. Using the nonparametric tests should account for the skew of the data when interpreting the p value to be significant. | |

* Out of curiosity I ran a count to identify if a specific treatment arm contained the higher values in the post CX3CL1 marker.
* Text

  Description automatically generated
* Graphical user interface, table

  Description automatically generated with medium confidence
* We can see the majority of the higher end values belong to the Trastuzumab arm which could be related to treatment effect which may explain the second peak but is beyond the scope of this question.
* The test for normality was repeated for TNFA using the same code aforementioned.

|  |  |
| --- | --- |
| Pre-Treatment Arm: | Post-Treatment Arm: |
|  |  |
|  |  |
| * The data is right skewed with majority of the data lying in the lower range of values. The first bar in the histogram is well above the predicted probability of finding an expression level of this amount. * We can see the tight skew of the data in the QQ plot beneath also as the data points digress from the line of normal distribution. * The hypothesis test also returned a significant p value(<0.05) so we can confirm the data is not normally distributed. | * The histogram is not normally distributed on visual appraisal. This may be multimodal data with 2/3 peaks observed in the data. * The QQ plot does generally trend with the line but we can see the deviations of the peaks along the way. * The hypothesis also returned a significant p value(<0.05) indicating that the data is not normally distributed. * The Wilcox test to compare the medians will be used again. This should account for the skewness and the non-normal distribution. |
| Shapiro p value: p<3.422e-08 | Shapiro p value: p<4.747e-09 |
| * A box plot was used to visualise the data first:      * The medians of the ranks do appear to have a slight difference. The hypothesis test will identify if this is significant. We can also see the outliers of the TNFA pre data similar to the tailing seen in the box plot. | |
| * Text    Description automatically generated * The hypothesis test returned a significant p value(<0.05) indicating that there is a significant difference in the median of the ranked data. * There is a statistically significant difference in the expression levels of TNFA pre and post data. | |

* The tests for normality were repeated for CCL20

|  |  |
| --- | --- |
| Pre-Treatment Arm: | Post-Treatment Arm: |
|  |  |
|  |  |
| * The histogram shows the data is right skewed. With values between 0 and 0.5 well above the expected probability. * The tailing can be seen on the QQ plot also as the data points deviate from the line of normal distribution. * The hypothesis test also indicates non normal distribution with a significant p value(<0.05). | * The histogram shows that the data is not normally distributed, it has 2 exaggerated peaks with a third smaller one. * The QQ plot shows the deviations of the points from the line of normal distribution and the tailing off to the right. * The hypothesis test also returned a significant p value(<0.5) confirming the data is not normally distributed. * The data was visualised before the hypothesis test was conducted. |
| Shapiro p value: p<2.2e-16 | Shapiro p value: p<7.344e-06 |
| * You could assume there is a statistical significant difference between the medians of the ranked groups based on this visual representation. You can also see how the range in the pre-treatment levels are narrow compared to the spread of the post treatment values. * The data is non parametric so the Wilcox test will be used to compare the median of the ranks. | |
| Graphical user interface  Description automatically generated with low confidence   * The p value is significant(<0.05) so we can confirm that there is a statistically significant difference between the medians of the groups. | |

* All the inflammatory markers except for CCL20 decreased after treatment which could indicate the treatments have an overall anti-inflammatory effect except in the case of CCL20.

**Q2. This particular treatment regimen is quite severe (the backbone is paclitaxel and cisplatin). Loss of appetite, gastrointestinal upset and weight loss are common side effects. The treating clinician will intervene to mitigate this. However, in some cases a severe nutritional deficit will occur. Is there a significant increase in the number of patient underweight post treatment (classed as a BMI of < 20)?**

* First I got a count of the pre and post BMI<20.
* A picture containing text

  Description automatically generated
* Table

  Description automatically generated
* This table tells us:
* 1: 65 patients were not considered underweight before or after treatment.
* 2: 16 patients were not considered underweight before treatment but were after treatment.
* 3: 2 patients were considered underweight before treatment and not after treatment.
* 4: 5 were underweight before and after treatment.
* I converted this data in a matrix using the following code:
* Text

  Description automatically generated with medium confidence
* I then ran the McNemar test on the data to test for significance of related groups.
* Graphical user interface, text

  Description automatically generated
* The p value was significant (<0.05) which means there is a significant increase in the number of patients with underweight BMI’s as 16 patients became underweight after treatment.

**Q3. Is there a statistically significant difference between the expression levels of the 4 immune cell markers over the course of the trial (3 time points, pre, during and post), and where does that difference lie?**

* To compare the data over the different time points I will use the Friedman test. First I had to organise the data in long form. I created 3 different data frames, one for each variable with the different time points stacked.
* Graphical user interface, text

  Description automatically generated
* I then conducted the Friedman test to test for significance between the time points followed by Conover’s test of pair wise comparison to tell us where the difference is
* Graphical user interface, text

  Description automatically generated
* This was repeated for all the variables

|  |  |
| --- | --- |
| Variable | Test |
| IL1B | Graphical user interface, text  Description automatically generated with medium confidence   * The Friedman test tells us that there is a significant difference between the median of the ranks as the p value<0.05. |
| Text  Description automatically generated   * The Conover test shows us where the difference lies within the data. * There is a significant difference in the median of the ranks between IL1B pre and during, and between IL1B pre and post. * This would indicate that treatment had an impact on the IL1B levels and it did not return to baseline after treatment. |
| CX3CL1 | Text  Description automatically generated   * The Friedman test tells us that there is a significant difference between the median of the ranks as the p value<0.05. |
| Text  Description automatically generated   * There is a significant difference in the median of the ranks for CX3CL1 pre and during, and after and during. * This could indicate that while on treatment the expression levels were affected but returned to baseline/pretest levels after treatment |
| TNFA | Graphical user interface, text  Description automatically generated   * The Friedman test tells us that there is a significant difference between the median of the ranks as the p value<0.05. |
| Text  Description automatically generated   * There is a significant difference between all time points. * This could indicate the TNFA levels continuously changed and did not return to pretest levels. |
| CCL20 | Text  Description automatically generated   * The Friedman test tells us that there is a significant difference between the median of the ranks as the p value<0.05. |
| Text  Description automatically generated   * This shows a significant difference between all the time points for the median of the ranks. * The expression levels clearly changed from the treatment and did not return to baseline after treatment. |