Predictors of Sexual Dysfunction among Breast Cancer Survivors

Client: Michelle Yang
Consultant: Wakeel Kasali

Abstract

Breast cancer survivors often experience hormonal changes, surgical shock, fatigue, physical effects, altered relationship dynamics, and age-related changes, which collectively contribute to sexual dysfunction and impact intimacy. This study explores how varying sexual functioning measures, including sexual satisfaction and arousal, respond to the influences of physical activity and age at diagnosis among the breast cancer survivors.

To explore the functional relationships between each measure and a predicting factor in this study, various statistical procedures may be employed. Ordinal or multinomial regression is recommended for predicting sexual functioning measures on an ordinal scale depending on how "Preferred not to answer" and "No response" are handled. Continuous outcome variables are modeled using linear regression, assuming the normality and homoscedasticity of deviations from the prediction equation. Diagnostics, including residual and Q-Q plots, are applied to validate model assumptions. Transformations before regression are decided based on scatterplots.

1. Introduction

Sexual dysfunction is a prevalent issue, affecting approximately 25% to 63% of women, and significantly impacts personal well-being and relationships. Despite the seriousness of the problem, a gap exists in understanding the association among sexual functioning measures, demographic, and behavioral factors, particularly among breast cancer survivors. This study aims to identify predictors of sexual dysfunction by examining the relationships between sexual functioning measures—such as *sexual satisfaction, pleasure, activity, arousal, lubrication, the Sexual Interest and Desire Inventory-Female (SIDI-F), the Female Sexual Distress Scale (FSDS), vaginal pain, and distress*—and demographic and behavioral factors, including *physical activity levels* and *age at diagnosis*. This study hypothesizes that lower *physical activity levels* are associated with reduced *sexual pleasure* and *overall functioning*. It also posits that older *age at diagnosis* is associated with lower *sexual satisfaction* and *activity* but may enhance *overall functioning*. The following sections describe the dataset, formulate the key statistical questions, and outline the recommended methods for analysis.

2. Data description and collection

The baseline data used in this study are obtained from 116 breast cancer survivors from British Columbia and Alberta who participated in a randomized controlled trial (RCT) in 2021. The parent dataset evaluates the effects of two interventions: an online mindfulness-based therapy and an online psychoeducational program. This report concerns the data collected before the interventions, with a focus on sexual dysfunction in breast cancer survivors. The online Qualtrics survey was used during the COVID-19 pandemic for participants to respond to a series of questions assessing demographic characteristics, lifestyle factors, and sexual functioning measures.

The explanatory variables are *age at diagnosis* (in years) and *physical activity* (measured as average hours per week). These are continuous variables with no missing values. The outcomes variables are (i) *sexual pleasure*, *(ii) SIDI-F*, (iii) *vaginal pain*, (iv) *FSDI*, (v) *sexual satisfaction*, (vi) *sexual activity*, (vii) *lubrication*, and (viii) *arousal*, collectively referred to as sexual functioning measures. They are measured under both continuous and ordinal scales. The response options for the outcome variables include "Not applicable" (a valid response indicating that the question does not apply), "Prefer not to answer" (a deliberate refusal to provide information despite relevance), and Likert-scale categories (for ordinal data). "No response" indicates missing value due to no answer to the item. The *SIDI-F* and *FSDS* are continuous scales, each derived from approximately 13 items, with item missing rates of 0.9% and 6.8%, respectively, due to "No response". The ordinal variables—*vaginal pain, sexual satisfaction, sexual pleasure (partnered), sexual arousal (partnered and solo), and sexual activity (partnered and solo)*—exhibit varying patterns of a blank: "No response" (0.9% to 3.4%), "Prefer not to answer" (0.9%), and "Not applicable" (up to 46.6%). The highest percentage of "Not applicable" was recorded 9.5% in *sexual satisfaction*, 46.6% in *sexual pleasure (partnered)*, 30.2% in *sexual arousal (solo)*, and 12.1% in *sexual activity (partnered)*.

3. Statistical Goals

The objective is to understand the relationship between x and y for the following (x, y) pairs: **(i.)** x_1 (*physical activity*) and y_1 (*sexual pleasure*); **(ii.)** x_1 and y_2 (*SIDI-F*); **(iii.)** x_1 and y_3 (*vaginal pain*); **(iv.)** x_1 and y_4 (*FSDI*); **(v.)** x_2 (*age at diagnosis*) and y_5 (*sexual satisfaction*); **(vi.)** x_2 and y_6 (*sexual activity*); **(vii.)** x_2 and y_2; **(viii.)** x_2 and y_3; **(ix.)** x_2 and y_4; **(x.)** x_2 and y_7; **(xi.)** x_2 and y_8 (*arousal*).

The main statistical objectives of the study can be summarized as follows:

1. How do *age at diagnosis* and *physical activity* predict different types of sexual dysfunction among breast cancer survivors?
2. What are the preliminary statistical steps required before conducting the analysis, and what are the follow-up statistical steps ?

4. Exploratory Data Analysis (EDA)

Using plots and tables, the relationships between variables have to be assessed. The study design includes items with various response categories and missing values, a bar chart therefore is appropriate for displaying the percentage distribution across different items/scales.

A bar chart can present the prevalence of missing data. A frequency or percentage table should be created for each item to ensure that the pooled values are calculated accurately. Use this information to construct the bar chart. The chart uses pooled values derived from the frequency or percentage table for all items/scales. Figure 1 shows the non-ordinal response rates present in various items and scales of the study dataset, providing a visual summary to assess their distributions.

For "Not applicable" responses, which indicate that the item is irrelevant to the participant (e.g., questions about *sexual activity* with partners for those without a partner), such cases are excluded entirely for the corresponding item. In contrast, for "No response" and "Prefer not to answer," where their rates are negligible, then listwise-deletion may be considered reasonably.

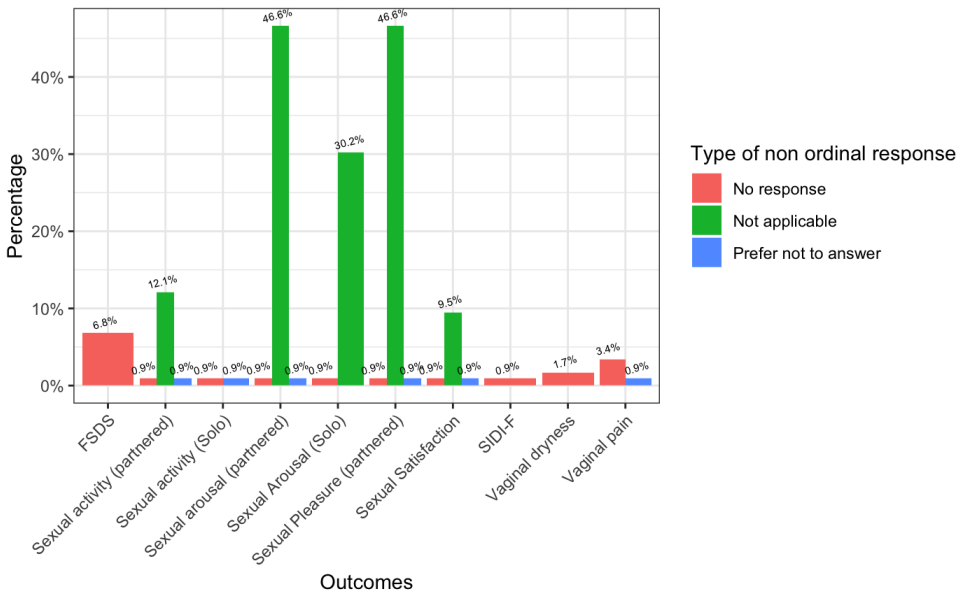

Figure 1: Bar chart of non ordinal rates for different items/scales in the study dataset. The y-axis represents rates as percentages, with corresponding percentages annotated on each bar

It would be reasonable to decompose the responses (non ordinal and ordinal) present in an item (see Appendix B) before finding the bivariate (two-item) relationships. When one variable tends to increase or decrease as another variable changes, but not necessarily at a constant rate, a monotonic relationship is present. This is evident in Appendix B, where the percentages of dissatisfaction (e.g., "Dissatisfied" and "Somewhat Dissatisfied") decrease with increased *physical activity*, but higher satisfaction levels (somewhat satisfied, satisfied) do not consistently rise with greater activity levels.

In SPSS, correlation tests are conducted to assess associations between a continuous outcome and an explanatory variable, one at a time, if there is no "Preferred not to answer" or "NA". For instance, Pearson correlation evaluates the linear association between continuous variables, such as y_2 and x_1. Spearman rank correlation, on the other hand, assesses monotonic associations between variables (ordinal and continuous), such as y_5 and x_1. A detailed instructional guide for performing Spearman correlation is available in (1).

A scatterplot, as shown in Appendix B, can be used to check whether to do data transformation or not before regression, recommending a square root transformation for the count variable x_1 (2).

5. Recommended statistical methods

Two statistical methods are recommended to address the two statistical questions and are described as follows:

5.1 Statistical method for categorical outcome

For an outcome variable, the "No response" and "NA" cases are omitted, so the sample size (< 116) depends on the outcome variable. Ordinal regression is an appropriate analytical approach when outcome variables have more than two categories/levels which are ordinal in nature, and the explanatory variables are continuous. This model estimates the probability of multiple ordered categories (e.g., no sexual activity, no desire/never aroused, little desire, moderate desire, and strong desire) in y_1 as a function of explanatory variables such as x_1. For ("No response"), listwise deletion is applied handle these cases.

When fitting an ordinal regression model, it is essential to check the assumption of **(constant slope)**. For example, in the case of (x_1, y_1), this assumption implies that the (x_1) has the same effect on the odds, regardless of the consecutive splits of the y_1 (e.g., no sexual activity versus all above, no sexual activity and

never aroused combined versus all above, etc.). If this assumption is violated, a partial proportional odds model (available in SAS) or a multinomial logistic regression model may be more appropriate. Moreover, this approach is only valid if EDA suggests a monotonic relationship. Otherwise, it may be necessary to merge categories in y_2 to create new categories, such as (Dissatisfied and Somewhat Dissatisfied), (Neutral), and (Somewhat Satisfied and Satisfied), to better meet model assumptions.

In contrast, multinomial regression models (See Appendix) do not require an ordinal structure in the outcome variable, as they treat the outcome as nominal. The two are available in R, SPSS and SAS. There might be a need to merge ordinal categories in order to get estimates if some categories are infrequent. Multinomial logit in multinomia model can handle the category of "Prefer not to answer".

5.2 Statistical method for continuous outcome

In the second part of the analysis, developing prediction models involving continuous outcomes and predictor variables (e.g., (x_1, y_2), (x_1, y_4), (x_2, y_2) and (x_2,y_4)) are tested using a suitable approach. For instance, fitting a Simple Linear Regression Model (SLRM) (see the statistical equation in the appendix) is appropriate for evaluating the relationship between x_1 and y_2 in (x_1, y_2).

After excluding cases with missing values, a Simple Linear Regression Model (SLRM) will be applied to examine the relationship between - say - physical activity (x_1) and sexual dysfunction scores (SIDI-F, y_2). Scatterplots is first used to confirm a monotone relationship between the variables, indicating that regression analysis is appropriate. Model assumptions can then be evaluated.

For the model involving (x_1, y_2), as defined in Equation 1, the residual plot would indicates that the residuals $e_1, \ldots, e_n$ are estimates of the error terms. If the regression model adequately describes the relationship between the variables, the residuals should behave like random noise, without systematic patterns in their relationship with the predictor variable (x_1). That will illustrate how sexual dysfunction (SIDI-F) varies with physical activity, assuming that unexplained variations remain consistent across all levels of activity.

Similarly, a Q–Q plot in the model diagnostics should display points that lie very close to or on the diagonal line. This pattern indicates that small errors occur frequently, large errors occur rarely, and errors are equally likely to be positive or negative, thereby satisfying the assumption of normality in the error terms. When at least one of the two assumptions (assessed through the residual plot and Q–Q plot) does not hold, a transformation of the outcome variable or adding the square the predictor (to suggest a polynomial relationship) may be considered. Consequently, other relationships (e.g., (x_1, y_4), (x_2, y_2), and (x_1, y_2)) using this approach should follow the same analytical pattern.

6.0 Conclusion

For sexual functioning measures on an ordinal scale, ordinal or multinomial regression is recommended to evaluate the influence of physical activity and age at diagnosis. Linear regression is recommended for continuous outcome variables. Scatterplots and two-way tables can help identify appropriate transformations of physical activity and age at diagnosis for use in regression models.

Further reading

1. Logistic Regression Models for Ordinal outcome Variables (3)

References

[1] L. Statistics. Spearman's rank-order correlation using spss statistics. [Online]. Available: https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php

[2] N. R. Draper and H. Smith, *Transformation of the Response Variable.* John Wiley Sons, Ltd, 1998, ch. 13, pp. 277–298. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118625590.ch13

[3] A. A. O'Connell, *Logistic regression models for ordinal response variables.* sage, 2006, vol. 146.

A. Appendices

A1. Model comparisons

A1.1 ORDINAL REGRESSION MODEL

Given the predictor x_1 (physical activity), the ordinal logistic regression with j of y_2 (sexual satisfaction) is defined such that such that we let Y represent the ordinal outcome variable with five categories. The cumulative log-odds of being in a category ℓ or below as a linear function of the explanatory variable:

$$\ln\left(\frac{\mathbb{P}(Y \leq \ell)}{\mathbb{P}(Y > \ell)}\right) = \zeta_\ell - \eta X, \quad \ell = 1, 2, \ldots, 4,$$

where ζ_ℓ represents the cumulative thresholds specific to each level ℓ, and η is the regression coefficient that quantifies the effect of physical activity on the log-odds of being in a lower category of sexual satisfaction.

A1.2 MULTINOMIAL LOGISTIC REGRESSION

Given the predictor x_1 (physical activity), the multinomial logistic regression models the probability of each category j of y_2 (sexual satisfaction) as follows:

$$p_j(x_1) := \mathbb{P}[y_2 = j \mid x_1] = \frac{e^{\beta_{0j} + \beta_{1j} x_1}}{1 + \sum_{\ell=1}^{4} e^{\beta_{0\ell} + \beta_{1\ell} x_1}}, \quad \text{for } j = 1, \ldots, 4, \tag{0.1}$$

(representing categories: Dissatisfied, Somewhat Dissatisfied, Neutral, Somewhat Satisfied, and with Satisfied as the reference level).

For the last level ($j = 5$, Satisfied):

$$p_5(x_1) := \mathbb{P}[y_2 = 5 \mid x_1] = \frac{1}{1 + \sum_{\ell=1}^{4} e^{\beta_{0\ell} + \beta_{1\ell} x_1}}. \tag{0.2}$$

Equations (0.1) and (0.2) imply that the probabilities across all categories sum to 1, i.e.,

$$\sum_{j=1}^{5} p_j(x_1) = 1. \tag{0.3}$$

There are $4 \times 2 = 8$ coefficients to estimate, as there are 4 non-reference categories and one predictor (x_1) plus an intercept for each.

The last level, $j = 5$ (Satisfied), serves as the *reference level*, a common convention though any category could be chosen. The multinomial logistic regression model has a useful interpretation in terms of logistic regressions. Taking the quotient of (0.1) and (0.2) gives:

$$\frac{p_j(x_1)}{p_5(x_1)} = e^{\beta_{0j} + \beta_{1j} x_1}, \quad \text{for } j = 1, \ldots, 4. \tag{0.4}$$

By applying a logarithm to both sides, gives:

$$\log\left(\frac{p_j(x_1)}{p_5(x_1)}\right) = \beta_{0j} + \beta_{1j} x_1, \quad \text{for } j = 1, \ldots, 4. \tag{0.5}$$

A2.1 LINEAR REGRESSION

Linear Regression Model is defined as:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where:

- Y: Continuous outcome variable (e.g., *SIDI-F)*,
- X: Continuous predictor variable (e.g., *physical activity*),
- β_0: Intercept, representing the expected value of Y when $X = 0$,
- β_1: Slope coefficient, indicating the change in Y for a one-unit change in X,
- ϵ: Error term, assumed to be normally distributed with a mean of 0 and constant variance.

$$\text{SIDI-F} = \beta_0 + \beta_1(\text{Physical Activity}) + \epsilon \tag{1}$$

The error term ϵ is assumed to follow a normal distribution with a mean of 0 and constant variance σ^2. This can be represented as:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

Transformation of x_1, y_2 that might be considered

$$\text{SIDI-F} = \beta_0 + \beta_1 * (\text{Physical activity}) + \beta_2 * (\text{Physical activity})^2 + \epsilon_i$$

B. Recommended descriptive statistical methods

Table 1: Distribution of physical activity levels and sexual satisfaction (ordinal responses) in percentage by row, based on real study data

Physical Activity (Hour)	Dissatisfied	Somewhat Dissatisfied	Neutral	Somewhat Satisfied	Satisfied	Total
Low (0–3)	23 (50.00)	17 (36.95)	2 (4.35)	2 (4.35)	2 (4.35)	46 (100)
Moderate (4–7)	19 (42.22)	17 (37.77)	3 (6.67)	3 (6.67)	3 (6.67)	45 (100)
Active (8–15)	4 (33.33)	5 (41.66)	2 (16.66)	1 (8.33)	0 (0)	12 (100)
Total	46	39	7	6	5	103

Table 2: Physical activity and non-ordinal sexual satisfaction responses based on real study data

Physical Activity (Hour)	No response	Not applicable	Prefer Not to Answer	Total
Low (0–3)	1	2	1	4
Moderate (4–7)	0	8	0	8
Active (8–15)	0	1	0	1
Total	1	11	1	13

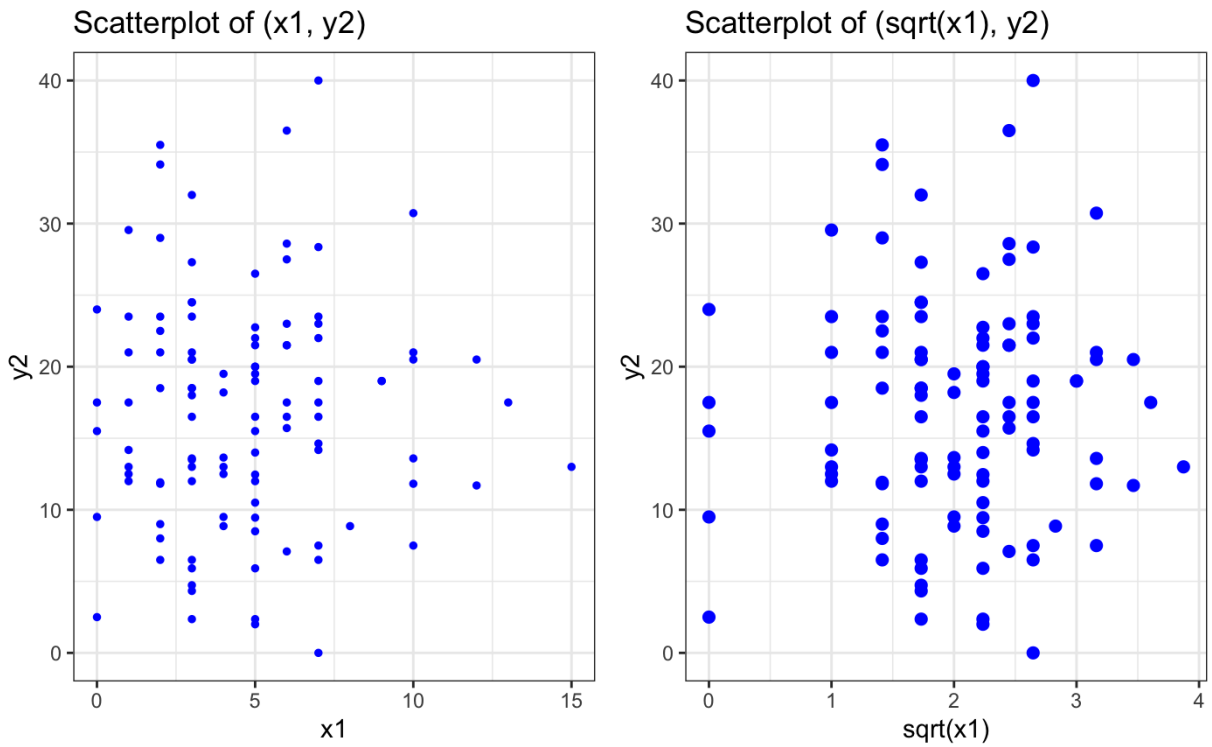

Figure 2: Scatterplots of x_1 and y_2: (x_1, y_2) on the left and $(\sqrt{x_1}, y_2)$ on the right (x_1 transformed) based on real study data. No association is observed.