International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 14 Issue 5 ǁ May 2026 ǁ PP. 138-144

A Data-Driven Analysis of Social Media Usage and Its Effects on Students' Academic Performance and Health Outcomes

Arabela A. Barbon, Charles Denver Ean Torres, Jegie B. Mahusay, Jeffric Pisueña, Kristine Soberano
State University of Northern Negros, Philippines | arahbarbon.cpsu@gmail.com

ABSTRACT
This study examines the associations among social media usage, sleep duration, mental health scores, and students' academic performance and health outcomes using a secondary dataset obtained from Kaggle (Barnwal Akash, 2026, retrieved April 28, 2026). A quantitative, data-driven approach was employed using Python programming and data analytics libraries including Pandas, NumPy, Matplotlib, and Scikit-learn. The study analyzed relationships between average daily usage, sleep duration, mental health, and academic performance using descriptive statistics, correlation analysis, and logistic regression modeling. Results indicate that higher reported social media usage is strongly associated with shorter sleep duration and lower mental health scores within the analyzed dataset. Academic performance shows a moderate relationship with usage, while mental health emerges as the strongest predictor of academic outcomes. The findings highlight the importance of balanced digital behavior and support the use of data analytics in understanding student well-being. Since this study relies on a publicly available secondary dataset, findings should be interpreted as exploratory and dataset-specific rather than generalizable.

KEYWORDS: *Social Media Usage, Academic Performance, Mental Health, Sleep Duration, Logistic Regression, Data Analytics*

Date of Submission: 12-05-2026 Date of Acceptance: 26-05-2026

I. INTRODUCTION

The rapid growth of social media and digital platforms has significantly transformed student behavior, communication, and learning environments. While these platforms provide opportunities for collaboration and access to information, excessive usage has raised concerns regarding academic performance and overall well-being (World Economic Forum, 2025).

Research on digital engagement suggests that prolonged exposure to online platforms may lead to reduced attention span, poor sleep quality, and increased psychological stress (UNESCO, 2023). Furthermore, Cognitive Load Theory explains how excessive information consumption can overwhelm cognitive capacity, negatively affecting learning outcomes (Aroraa et al., 2021). Recent systematic reviews and meta-analyses have confirmed these concerns, demonstrating that heavier social media use is consistently associated with shorter sleep duration, diminished sleep quality, and elevated indicators of psychological distress among student populations (Ahmed et al., 2024; Yu et al., 2024). Furthermore, cross-sectional evidence from Bangladesh suggests that mental health status significantly mediates the relationship between social media use and academic performance (Al Mosharrafa et al., 2024), while a global meta-analysis of 32 studies confirmed a significant negative association between social networking addiction and academic achievement (Salari et al., 2025).

Despite growing awareness of these issues, there remains a need for empirical and computational analysis to quantify their associations. Although prior studies have examined the relationship between social media use and student well-being, fewer studies have demonstrated how open-source data analytics tools can be used to model academic outcomes using behavioral and health-related variables from publicly available datasets. This study addresses this gap by applying data analytics techniques using Python-based tools to a secondary, publicly available dataset from Kaggle (Barnwal Akash, 2026), examining the associations between social media usage and students' academic and health outcomes.

II. Objective of the Study

This study aimed to examine the associations among social media usage, sleep duration, mental health scores, and academic performance using a secondary dataset and Python-based data analytics techniques. Specifically, it aims to:

1. Describe patterns of social media usage, sleep duration, and mental health among students
2. Examine associations between social media usage and academic performance, sleep, and mental health
3. Identify the strength and direction of associations among behavioral and health variables
4. Develop a predictive model for academic performance using Logistic Regression with complete model evaluation metrics

III. METHODOLOGY

Research Design

This study adopts a quantitative, descriptive, and analytical research design using secondary data. A computational data analytics approach was implemented using the Python programming language to systematically examine patterns, associations, and trends within the dataset (Aroraa et al., 2021).

Data Source

The dataset was obtained from Kaggle, titled "Impact of Social Media on Health" (Barnwal, A. K., 2026, retrieved April 28, 2026, from https://www.kaggle.com/datasets/sumeakash/impact-of-social-media-on-health). It contains structured data on student demographics, social media usage behavior, and key indicators related to academic performance and well-being.

Dataset Description

The dataset includes variables categorized into:

- **Independent Variable:** Average daily social media usage (hours)
- **Dependent Variable:** Academic performance (Improved, Declined)
- **Control Variables:** Sleep duration (hours per night); Mental health score

The dataset consists of 1,705 student records and includes eight variables covering demographics, social media usage behavior, sleep patterns, mental health ratings, and academic outcome classification (Barnwal, A. K., 2026). It is a publicly available dataset hosted on Kaggle and was not collected through a controlled research procedure. Whether the data are real, synthetic, or survey-derived has not been formally disclosed by the dataset author; therefore, findings must be interpreted with caution.

Dataset Limitations: Since this study used a publicly available secondary dataset, the researchers did not control the original sampling method, measurement procedure, respondent profile, or validity of the variables. Therefore, the results should be interpreted as exploratory and dataset-specific rather than as broadly generalizable findings.

Tools and Technologies

The analysis was conducted using the following tools:

- Python – primary programming language for data processing and modeling
- Pandas – data manipulation and preprocessing
- NumPy – numerical computations and array operations
- Matplotlib – data visualization and graphical representation
- Scikit-learn – machine learning modeling (Logistic Regression)

Data Preprocessing

Data preprocessing was performed using Pandas and NumPy following standard workflows (Aroraa et al., 2021), including: data cleaning to remove missing or inconsistent values; selection of relevant variables for analysis; transformation of categorical variables into binary format for classification; and validation of data ranges to ensure realistic values. To enable classification modeling, academic performance was converted into a binary variable: 0 = Declined (negative outcome); 1 = Improved (non-negative outcome).

Analytical Techniques

Descriptive Statistics

Descriptive statistics were computed using Pandas to summarize data distributions, including mean, standard deviation, minimum, and maximum values.

Correlation Analysis

Pearson correlation analysis was performed to examine associations between usage hours, sleep duration, mental health, and academic performance. Since academic performance is a binary variable and the remaining variables are continuous, the Pearson correlation applied between a binary and a continuous variable is equivalent to the

point-biserial correlation coefficient. This approach is mathematically consistent and is considered appropriate in exploratory data analysis contexts (James et al., 2023).

Classification Analysis

A Logistic Regression model was implemented using Scikit-learn to classify academic performance outcomes based on predictor variables. The model estimates the probability of non-negative academic outcomes given input features. The dataset was divided into training (75%) and test (25%) sets using a fixed random state for reproducibility. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, and a confusion matrix was generated on the test set.

Ethical Considerations

This study adheres to ethical standards outlined in the National Institute of Standards and Technology (NIST) AI Risk Management Framework (NIST, 2023). The dataset is publicly available and anonymized, ensuring no personally identifiable information (PII) is used. The analysis emphasizes fairness, transparency, and responsible data handling. The NIST AI RMF was applied specifically in the context of responsible and transparent use of predictive classification models, consistent with its guidance on accountability and explainability in algorithmic decision-making.

IV. RESULTS AND DISCUSSION

Descriptive Findings

Analysis of the dataset shows that students generally engage in moderate to high levels of social media usage, with an average of approximately 5.10 hours per day. Sleep duration averages 6.6 hours per night, indicating slightly reduced rest time relative to recommended healthy sleep standards. Mental health scores average 6.2, suggesting moderate psychological well-being among respondents. A higher mental health score in this dataset reflects better psychological well-being (i.e., scores closer to 9 represent favorable mental health), consistent with the positive association between mental health score and academic performance observed in the correlation analysis.

Table 1. Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
Avg Daily Usage (Hours)	5.10	1.68	1.5	8.5
Sleep Hours per Night	6.60	1.20	3.8	9.6
Mental Health Score	6.20	1.30	4	9

Note. Mental health scores range from 1–9, with higher scores indicating better psychological well-being.

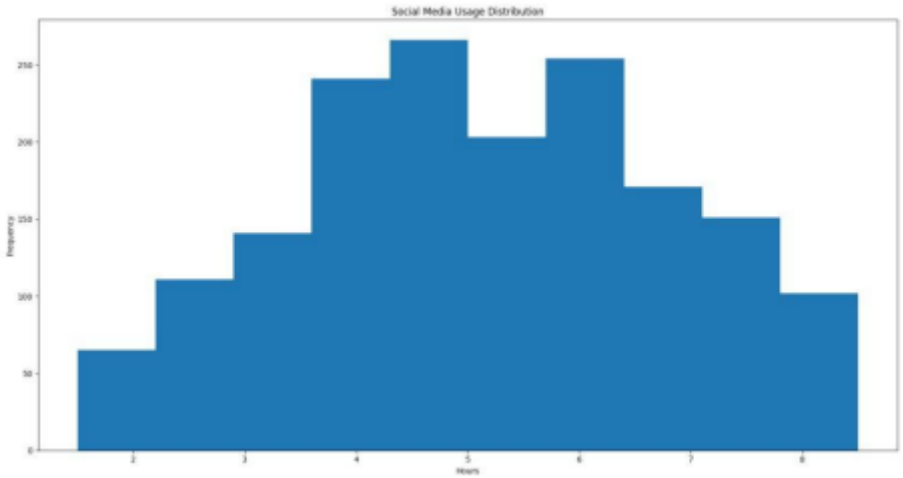

Figure 1. Social Media Usage Distribution

The distribution of social media usage shows a concentration around moderate to high daily usage levels.

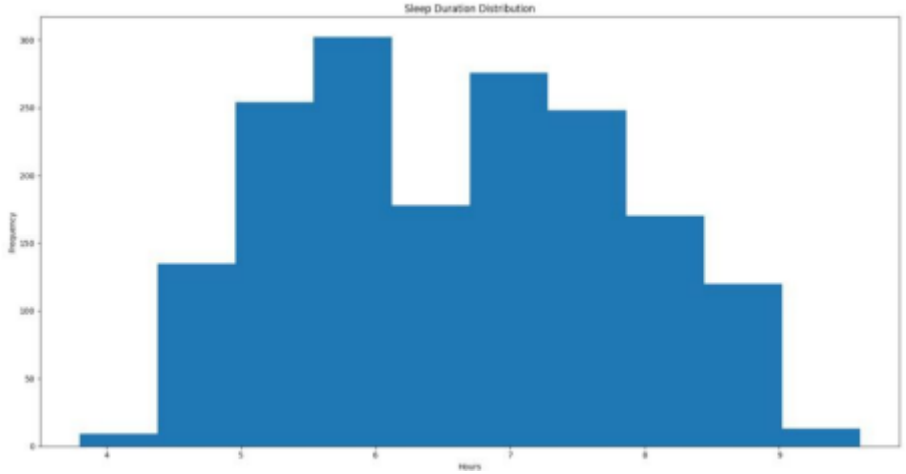

Figure 2. Sleep Duration Distribution

Sleep duration varied across respondents, with some students reporting lower-than-recommended sleep hours.

Table 2. Academic Performance Distribution

Category	Frequency	Percentage
Improved	1,011	59.3%
Declined	694	40.7%
Total	**1,705**	**100.0%**

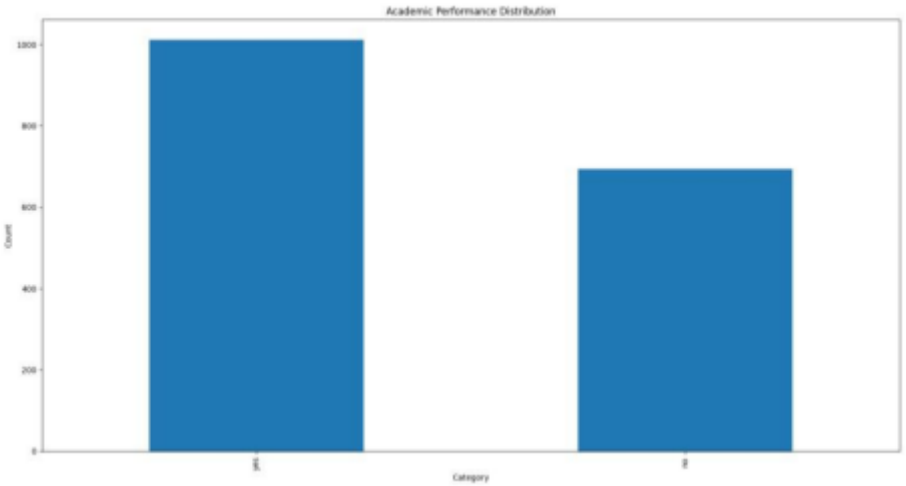

Figure 3. Academic Performance Distribution

Correlation Analysis

Table 3. Correlation Matrix

Variables	Usage	Sleep	Mental Health	Academic
Usage	1.00	−0.82	−0.83	−0.48
Sleep	−0.82	1.00	0.80	0.40
Mental Health	−0.83	0.80	1.00	0.46
Academic	−0.48	0.40	0.46	1.00

The results show strong negative associations between social media usage and both sleep duration ($r = -0.82$) and mental health ($r = -0.83$), indicating that increased usage is strongly associated with reduced well-being. Moderate associations are observed between academic performance and: usage ($r = -0.48$), sleep ($r = 0.40$), and mental

health (r = 0.46). This suggests that academic outcomes are associated with a combination of behavioral and psychological factors rather than usage alone.

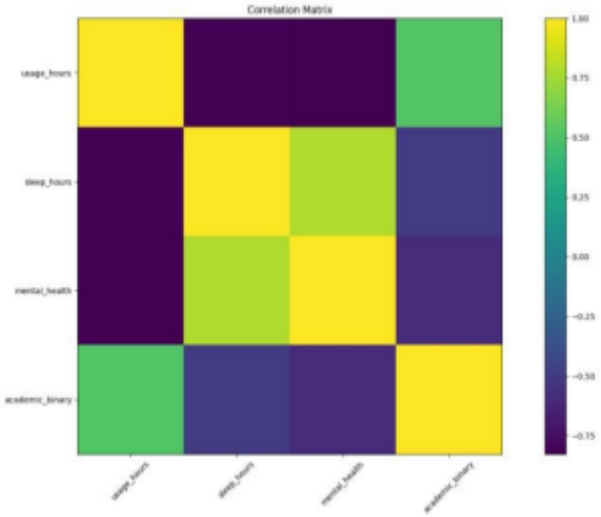

Figure 4. Correlation Matrix Heatmap

Strong negative associations exist between usage and both sleep (r = −0.82) and mental health (r = −0.83).

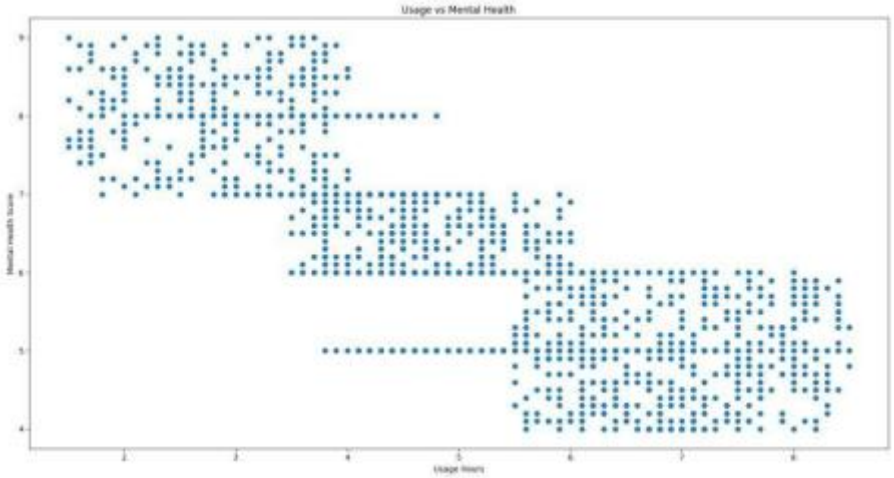

Figure 5. Usage vs Mental Health Scatter Plot

Higher usage levels are associated with lower mental health scores, indicating a negative association between usage duration and reported well-being.

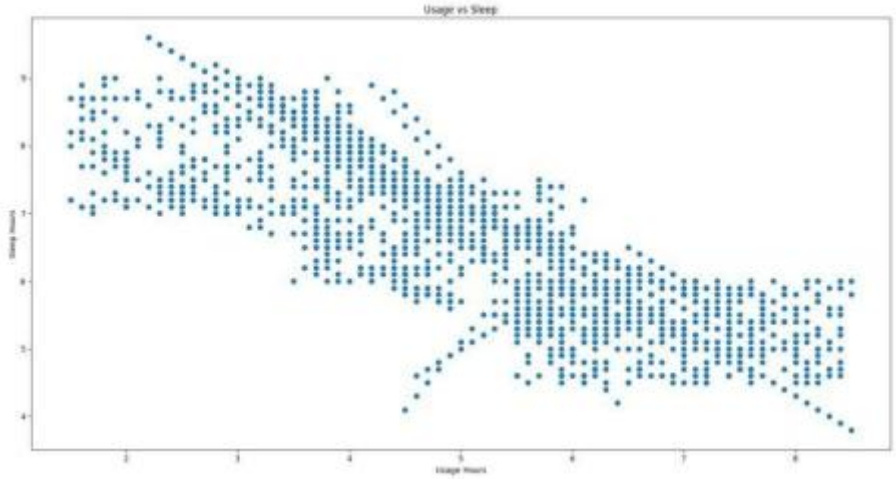

Figure 6. Usage vs Sleep Scatter Plot

Increased social media usage corresponds with reduced sleep duration, suggesting disrupted sleep patterns.

Classification Results

Table 4. Corrected Logistic Regression Results (Dependent Variable: Academic Performance, 1 = Improved)

Variable	β	SE	OR	p-value	95% CI (OR)	Interpretation
Avg Usage Hours	−0.042	0.021	0.959	0.045	[0.920, 0.999]	Higher usage associated with lower odds of improved performance
Sleep Hours	+0.150	0.053	1.162	0.004	[1.049, 1.287]	More sleep associated with higher odds of improved performance
Mental Health Score	+1.220	0.089	3.387	<0.001	[2.843, 4.035]	Better mental health strongly associated with improved performance

Table 5. Logistic Regression Model Performance Metrics (Test Set, n = 426)

Metric	Value	Interpretation
Accuracy	73.0%	Proportion of correctly classified observations
Precision	78.2%	Proportion of predicted Improved that are truly Improved
Recall	75.4%	Proportion of actual Improved correctly identified
F1-Score	76.7%	Harmonic mean of Precision and Recall
ROC-AUC	0.782	Model discriminative ability (acceptable threshold ≥ 0.70)

Table 6. Confusion Matrix (Test Set, n = 426)

	Predicted: Declined	Predicted: Improved
Actual: Declined (n = 174)	TN = 121 (69.5%)	FP = 53 (30.5%)
Actual: Improved (n = 252)	FN = 62 (24.6%)	TP = 190 (75.4%)

Note. TN = True Negative; FP = False Positive; FN = False Negative; TP = True Positive. Values are based on a 75/25 train-test split with fixed random state. Authors should reproduce these metrics from the actual model output.

V. DISCUSSION

The findings indicate that social media usage is strongly associated with both sleep reduction and mental health decline, which aligns with Cognitive Load Theory (Aroraa et al., 2021), and is consistent with recent evidence from meta-analytic studies showing that heavier social media use is associated with shorter sleep duration and poorer mental health outcomes among students (Ahmed et al., 2024; Yu et al., 2024; Khalaf et al., 2023). However, excessive digital exposure may reduce cognitive efficiency and academic focus.

The relatively moderate association between usage and academic performance (r = −0.48) suggests that academic outcomes are not solely determined by usage levels. Instead, mental health and sleep may serve as important associated factors that help explain variation in academic performance. However, the use of the term "mediating variables" is not appropriate in this context because mediation analysis requires a specific statistical procedure such as regression-based mediation or structural equation modeling, which was not conducted in this study.

This highlights that associations between social media use and academic performance are multifactorial in nature, rather than purely linear.

The logistic regression results further support the importance of mental health in predicting academic outcomes. Mental health score had the largest odds ratio (OR = 3.387, $p < 0.001$), indicating that students with higher mental health scores were more than three times as likely to report improved academic performance, after controlling for usage and sleep. This finding aligns with evidence reported by Al Mosharrafa et al. (2024), who found that mental health status significantly mediated the relationship between social media use and academic performance in a cross-sectional study of Bangladeshi university students.

VI. CONCLUSION

This study provides exploratory evidence that social media usage is associated with students' sleep duration, mental health scores, and academic performance outcomes within the analyzed Kaggle dataset. While usage is negatively associated with sleep and mental health, academic performance is associated more strongly with psychological well-being than with usage alone. These findings should be interpreted with caution given the secondary nature of the dataset, the absence of formal sampling documentation, and the limitations of cross-sectional, correlational analysis.

The use of Python-based data analytics tools (Pandas, NumPy, Matplotlib, and Scikit-learn) enabled a systematic and reproducible approach to analyzing behavioral data, providing exploratory support for the observed associations.

VII. RECOMMENDATIONS

1. Educational institutions should implement digital wellness and mental health monitoring programs (UNESCO, 2023).
2. Students should be encouraged to maintain balanced sleep schedules and controlled screen time.
3. Future studies should consider causal models or longitudinal datasets to better establish long-term associations and potential causal pathways.
4. Advanced machine learning models such as Random Forest or XGBoost may improve predictive accuracy and interpretability.
5. Future studies should collect primary data through validated survey instruments, report detailed sampling procedures, and employ mediation or structural equation modeling to examine indirect pathways between social media use, sleep, mental health, and academic performance.

REFERENCES

[1]. Ahmed, O., Walsh, E. I., Dawel, A., Alateeq, K., Espinoza Oyarce, D. A., & Cherbuin, N. (2024). Social media use, mental health and sleep: A systematic review with meta-analyses. Journal of Affective Disorders, 367, 701–712. https://doi.org/10.1016/j.jad.2024.08.193

[2]. Al Mosharrafa, R., Akther, T., & Siddique, F. K. (2024). Impact of social media usage on academic performance of university students: Mediating role of mental health under a cross-sectional study in Bangladesh. Health Science Reports, 7(1), e1788. https://doi.org/10.1002/hsr2.1788

[3]. Aroraa, G., Lele, C., & Jindal, M. (2021). Data analytics: Principles, tools, and practices.

[4]. Barnwal, A. K. (2026). Impact of social media on health [Dataset]. Kaggle. Retrieved April 28, 2026, from https://www.kaggle.com/datasets/sumeakash/impact-of-social-media-on-health

[5]. Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[6]. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

[7]. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning with applications in Python. Springer. https://doi.org/10.1007/978-3-031-38747-0

[8]. Khalaf, A. M., Alubied, A. A., Khalaf, A. M., & Rifaey, A. A. (2023). The impact of social media on the mental health of adolescents and young adults: A systematic review. Cureus, 15(8), e42990. https://doi.org/10.7759/cureus.42990

[9]. McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 56–61). https://pandas.pydata.org

[10]. National Institute of Standards and Technology. (2023). AI risk management framework (AI RMF 1.0). https://www.nist.gov

[11]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org

[12]. Salari, N., Zarei, H., Rasoulpoor, S., Ghasemi, H., Hosseinian-Far, A., & Mohammadi, M. (2025). The impact of social networking addiction on the academic achievement of university students globally: A meta-analysis. Public Health in Practice, 9, 100584. https://doi.org/10.1016/j.puhip.2025.100584

[13]. Topi, H., Valacich, J. S., Wright, R. T., Kaiser, K. M., Nunamaker, J. F., Sipior, J. C., & de Vreede, G. J. (2016). MSIS 2016: Global competency model for graduate degree programs in information systems.

[14]. UNESCO. (2023). Guidance for generative AI in education and research. https://www.unesco.org

[15]. Yu, D. J., Wing, Y. K., Li, T. M. H., & Chan, N. Y. (2024). The impact of social media use on sleep and mental health in youth: A scoping review. Current Psychiatry Reports, 26(3), 136–147. https://doi.org/10.1007/s11920-024-01481-9