DATA TRANSFER AGREEMENT

For PLCO Data

Please complete the information below:	
CDAS PROJECT NUMBER:	PLCO-1742
PROJECT TITLE:	Exploring Multimodal Factors Influencing Breast Cancer Risk: A Data-Driven Approach
RECIPIENT:	Universitat de Barcelona
RECIPIENT LEAD INVESTIGATOR:	Oliver Díaz

The National Cancer Institute (NCI) and the RECIPIENT hereby enter into this Agreement for the transfer of data collected in the course of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (DATA) to RECIPIENT through NCI's Cancer Data Access System (CDAS). Collectively or individually, NCI and RECIPIENT shall also be referred to as Parties or Party. This Agreement is effective as of the date of the last signature below (Effective Date).

In consideration of NCI providing DATA to RECIPIENT, RECIPIENT hereby agrees to the following terms and conditions:

1. DATA WILL NOT BE USED TO TREAT OR DIAGNOSE HUMAN SUBJECTS. RECIPIENT will use DATA in compliance with all applicable local, state, and/or federal laws and regulations, including but not limited to those for the protection of human subjects.

2. RECIPIENT must not use DATA for any study other than the approved Research Plan, attached as **Attachment 1**, unless RECIPIENT obtains the written consent of NCI by way of a new approved application through CDAS or by written and signed amendment to this Agreement. RECIPIENT grants NCI the right to publicly disclose the Research Plan, including titles, summaries or any other information contained therein as well as the names and contact information for the investigators conducting the research.

3. The DATA will be used solely by RECIPIENT LEAD INVESTIGATOR and RECIPIENT's faculty, employees, fellows, students, and agents that have a need to use, or provide a service in respect of, the DATA in connection with the Research Plan and whose obligations for using the DATA are consistent with the terms of this Agreement.

4. The DATA will not be further distributed to others without NCI's written consent. The RECIPIENT shall refer any request for the DATA to NCI.

5. Personally identifiable information will not be provided. If DATA being provided are coded, RECIPIENT will not request the key to the code. RECIPIENT must not attempt to learn the identity of or to contact the human subjects from which DATA were obtained, their physicians, or the collection sites for DATA. In the event that personally identifiable information is inadvertently transferred, RECIPIENT agrees to immediately destroy the personally identifiable information and report the circumstances to NCI. The DATA may be protected by the Federal Privacy Act and/or a Certificate of Confidentiality.

6. DATA are the property of NCI and are made available as a service to the research community. RECIPIENT will not claim, infer, or imply ownership of DATA or any endorsement of RECIPIENT'S activities or products by the U.S. Government, DHHS, NIH, NCI, or NCI employees.

7. NCI MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF DATA WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS. Unless prohibited by law, RECIPIENT assumes all liability for claims for damages against it by third parties which may arise from its use, storage, or disposal of DATA.

8. RECIPIENT will acknowledge NCI as the source of DATA in all publications and presentations by including language substantially similar to the following: "The authors thank the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial". Each publication and presentation should reference the CDAS Project Number.

9. RECIPIENT must submit a description of each publication resulting from its use of DATA to the following website: https://cdas.cancer.gov/projects/plco/1742/PLCO-1742/discussion/. RECIPIENT agrees that NCI may publicly disclose this description.

10. This Agreement shall be in effect for five (5) years from the Effective Date. At the end of these five (5) years, if RECIPIENT is still using DATA for the approved Research Plan, RECIPIENT may seek an amendment to extend the term of this Agreement. This Agreement may be terminated by either Party for any reason by providing written notice to the other Party at least thirty (30) days prior to the desired termination date. Upon expiration or earlier termination of this Agreement or if RECIPIENT's use of DATA is complete, RECIPIENT must destroy DATA and upon NCI's request, confirm in writing as to such destruction. The RECIPIENT may retain one (1) copy of the DATA to the extent necessary to comply with the records retention requirements under any law, and for the purposes of research integrity and verification.

SIGNATURES BEGIN ON THE NEXT PAGE

ACCEPTED AND AGREED

FOR THE RECIPIENT (Universitat de Barcelona)

_____ _____

(Authorized Signatory for Recipient) Date

Printed Name _____

Title _____

Address _____

_____

_____

Read and Acknowledged by Recipient Lead Investigator:

Signature _____ _____

Name Oliver Díaz Date

Address _____

_____

_____

Attachment 1

The Cancer Data Access System

PLCO-1742 Research Plan for Universitat de Barcelona

Requestor

Name: Oliver Díaz

Institution: Universitat de Barcelona

Email: oliver.diaz@ub.edu

Recipient Lead Investigator

Name: Oliver Díaz

Institution: Universitat de Barcelona

Degrees: Ph.D.

Position: Associate Professor

Country: Spain

City: Barcelona

Email: oliver.diaz@ub.edu

Project Information

Title:

Exploring Multimodal Factors Influencing Breast Cancer Risk: A Data-Driven Approach

Summary:

The proposed research aims to investigate the relationship between various patient characteristics, including age, ethnicity, medication, lifestyle factors, and their correlation with breast cancer risk using the PCLO Breast dataset. By employing advanced data science techniques and machine learning algorithms, this study seeks to uncover new patterns and potential biomarkers that could enhance our understanding of breast cancer development and improve risk assessment.

Breast cancer is a complex and multifactorial disease, with numerous factors contributing to its incidence and progression. While significant progress has been made in early detection and treatment, there is a pressing need to further elucidate the underlying mechanisms and identify novel risk factors that can aid in the development of more personalized and effective preventive strategies.

The PCLO Breast dataset provides a rich source of information that has the potential to unveil new insights into the complex interplay between patient demographics, clinical characteristics, and breast cancer risk. By leveraging this comprehensive dataset, the proposed research will explore the correlation between a wide range of patient-level variables, such as age, ethnicity, medication history, lifestyle factors (e.g., physical activity, diet, smoking), and the risk of developing breast cancer.

Through the application of advanced data science and machine learning techniques, the research team aims to identify patterns, associations, and potential biomarkers that can contribute to a more nuanced understanding of breast cancer risk. This knowledge can lead to the development of improved risk assessment tools, personalized screening and intervention strategies, and ultimately, enhanced patient outcomes.

The findings of this study will not only advance the scientific understanding of breast cancer but also have the potential to inform clinical practice and guide the development of targeted prevention and early detection programs. By uncovering new insights into the multifaceted nature of breast cancer risk, this project can pave the way for more effective and tailored approaches to breast cancer management and ultimately improve the quality of life for individuals at risk.

Aims:

1.Comprehensive Data Analysis:
-Conduct a thorough exploration and preprocessing of the PCLO Breast dataset to ensure data quality and integrity.
-Perform extensive data visualizations and statistical analyses to identify significant associations between patient characteristics and breast cancer risk.
-Assess the interplay between demographic factors, clinical variables, and lifestyle-related attributes in the context of breast cancer development.

2.Machine Learning Model Development:
-Develop and train advanced machine learning models, such as logistic regression, decision trees, random forests, and neural networks, to predict breast cancer risk based on the identified risk factors.
-Evaluate the performance of these models using appropriate metrics, including accuracy, precision, recall, and F1-score, to ensure robust and reliable risk assessment.
-Explore the use of ensemble techniques and feature engineering to further enhance the predictive capabilities of the models.

3.Biomarker Discovery:

-Employ feature importance and selection techniques to identify the most influential variables contributing to breast cancer risk.
-Investigate the potential of these variables as novel biomarkers that can be used for early detection, risk stratification, and targeted interventions.
-Validate the identified biomarkers through additional statistical analyses and, if feasible, through external validation using independent datasets.

4.Model Interpretation and Explainability:
-Implement interpretable machine learning methods, such as SHAP (Shapley Additive Explanations), to provide insights into the underlying relationships between the input variables and the predicted breast cancer risk.
-Communicate the model's decision-making process in a transparent and understandable manner, enabling better clinical interpretation and facilitating the integration of the developed models into healthcare decision-making processes.

5.Translational Research and Clinical Implications:
-Explore the potential clinical applications of the developed risk assessment models and identified biomarkers.