Datasheet: Predictors of Multiple Sclerosis Disease

Jeffrey Kong, Noah Lee, Jacob Ventura

1. Dataset Information

This datasheet details the dataset created to investigate predictors of multiple sclerosis disease. It includes patient demographics, clinical history, and diagnostic test results.

2. Dataset Column Descriptions

Column Name	Description
ID	Patient identifier (int)
Age	Age of the patient (in years)
Schooling	Time the patient spent in school (in years)
Gender	1 = male, 2 = female
Breastfeeding	1 = yes, 2 = no, 3 = unknown
Varicella	1 = positive, 2 = negative, 3 = unknown
Initial_Symptoms	1 = visual, 2 = sensory, 3 = motor, 4 = other, 5 = visual and sensory, 6 = visual and motor, 7 = visual and others, 8 = sensory and motor, 9 = sensory and other, 10 = motor and other, 11 = Visual, sensory and motor, 12 = visual, sensory and other, 13 = Visual, motor and other, 14 = Sensory, motor and other, 15 = visual, sensory ,motor and other
Mono_or_Polysymptomatic	1 = monosymptomatic, 2 = polysymptomatic, 3 = unknown
Oligoclonal_Bands	0 = negative, 1 = positive, 2 = unknown
LLSSEP	0 = negative, 1 = positive
ULSSEP	0 = negative, 1 = positive
VEP	0 = negative, 1 = positive
BAEP	0 = negative, 1 = positive
Periventricular_MRI	0 = negative, 1 = positive
Cortical_MRI	0 = negative, 1 = positive
Infratentorial_MRI	0 = negative, 1 = positive
Spinal_Cord_MRI	0 = negative, 1 = positive
initial_EDSS	Expanded Disability Status Scale at diagnosis
final_EDSS	Expanded Disability Status Scale at follow-up
Group	1 = CDMS, 2 = non-CDMS

3. Additional Dataset Context

This dataset was collected to better understand potential predictors and diagnostic markers for multiple sclerosis disease progression. The dataset includes data on clinical demographics, diagnostic tests, and patient-reported symptoms.

4. Questions

Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to identify predictive factors for the conversion of Clinically Isolated Syndrome (CIS) to Clinically Definite Multiple Sclerosis (CDMS). It addresses a gap in understanding CIS progression in Mexican mestizo patients.

2. Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

The dataset was created by Researchers Víctor Chavarria, Guillermo Espinosa-Ramírez, Julio Sotelo, José Flores-Rivera, Omar Anguiano, Ana Campos Hernández, Edgar Daniel Guzmán-Ríos, Aleli Salazar, Graciela Ordoñez, and Benjamin Pineda. The dataset was created on behalf of the National Institute of Neurology and Neurosurgury (NINN) in Mexico City, Mexico.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The dataset was funded by Consejo Nacional de Ciencia y Tecnología (CONACYT) under grant number Salud-2012-01-181031.

4. Any other comments?

The dataset contributes to understanding the predictors of CIS to CDMS conversion, particularly emphasizing demographic and clinical aspects in Mexican patients.

Composition

5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances represent patients diagnosed with Clinically Isolated Syndrome (CIS) and their clinical, demographic, and diagnostic test data.

6. How many instances are there in total?

There are 273 total instances in the dataset.

7. Does the dataset contain all possible instances, or is it a sample? If it is a sample, what is the larger set? Is the sample representative?

Appears to be a sample of CIS patients who presented at the National Institute of Neurology and Neurosurgery (NINN) in Mexico City between 2006 and 2010. Representativeness is not explicitly discussed.

8. What data does each instance consist of (e.g., raw data or features)?

Each instance consists of demographic (e.g., age, gender), clinical (e.g., initial symptoms, MRI results), and diagnostic test results (e.g., BAEP, VEP). See **Dataset Column Descriptions** above.

9. Is there a label or target associated with each instance?

Yes, the label is **Group**, where 1 indicates CDMS (Clinically Definite Multiple Sclerosis) and 2 indicates non-CDMS.

10. Is any information missing from individual instances?

Yes, some data is missing or marked as "unknown" for variables like breastfeeding, varicella history, and some MRI results. There are also NA values for columns Initial_EDSS and Final_EDSS, which appear missing when CDMS=1.

11. Are relationships between individual instances made explicit?

No relationships between instances are described.

12. Are there recommended data splits (e.g., training, validation, testing)?

No recommended splits are provided.

13. Are there any errors, sources of noise, or redundancies in the dataset?

Some variables have missing or unknown values that could introduce noise in statistical models.

14. Is the dataset self-contained, or does it link to external resources?

The dataset is self-contained.

15. Does the dataset contain data that might be considered confidential?

It does not appear to include identifiable information other than patient ID.

16. Does the dataset contain data that might cause offense or anxiety?

The dataset deals with medical conditions, which may cause anxiety to those affected by similar conditions. No other data was identified to possibly cause offense or anxiety.

17. Does the dataset identify any subpopulations?

Yes, it focuses on Mexican mestizo patients with CIS.

18. Is it possible to identify individuals directly or indirectly?

No, the dataset does not include personally identifiable information.

19. Does the dataset contain data that might be considered sensitive?

Yes, medical data is considered sensitive.

20. Any other comments?

The dataset's focus on a specific population provides valuable insights but may limit generalizability to other populations.

Collection Process

21. How was the data associated with each instance acquired?

Data was collected through clinical evaluations, diagnostic tests, and follow-ups of CIS patients at the National Institute of Neurology and Neurosurgery in Mexico City.

22. What mechanisms or procedures were used to collect the data?

Clinical examinations, diagnostic imaging, and laboratory tests were used to collect the data.

23. If the dataset is a sample, what was the sampling strategy?

The sampling strategy is not explicitly described.

24. Who was involved in the data collection process, and how were they compensated?

The study's authors likely collected the data; details on compensation are not provided.

25. Over what timeframe was the data collected?

The data was collected between 2006 and 2010.

26. Were any ethical review processes conducted?

Not explicitly mentioned.

27. Did you collect the data directly from individuals or via third parties?

Data was collected directly from individuals during their clinical visits. This data was published and made downloadable from Kaggle.

28. Were the individuals notified about the data collection?

Not explicitly mentioned.

29. Did the individuals consent to the collection and use of their data?

Not explicitly mentioned.

30. If consent was obtained, were individuals provided a mechanism to revoke it?

Not mentioned.

31. Has an analysis of the dataset's impact on data subjects been conducted?

Not mentioned.

32. Any other comments?

The dataset is valuable for understanding CIS progression in a specific demographic.

Preprocessing, Cleaning, and Labeling

33. Was any preprocessing, cleaning, or labeling of the data done?

The description does not provide preprocessing or cleaning details, but "unknown" values are present. Preprocessing is needed for future model building.

34. Was the raw data saved in addition to the processed data?

Not mentioned.

35. Is the software used for preprocessing available?

Not mentioned.

36. Any other comments?

Unknown values could impact analysis and model performance, so considerations will need to be made on what to do with missing values or whether to include columns like Initial_EDSS.

Uses

37. Has the dataset been used for any tasks already?

The cited study used the dataset to identify predictors of conversion from CIS to CDMS.

38. Is there a repository linking to papers or systems using the dataset?

Yes, the dataset is linked to the study published in the Archives of Medical Research.

39. What other tasks could the dataset be used for?

Tasks the dataset could be used for include:

- Predictive modeling for CIS to CDMS conversion.
- Exploratory analysis of CIS symptoms and demographics.
- Understanding relationships between clinical and diagnostic markers.

40. Is there anything about the dataset's composition or collection that might impact future uses?

The focus on a single demographic group (Mexican mestizo patients) may limit generalizability.

41. Are there tasks for which the dataset should not be used?

The dataset should not be used for tasks unrelated to its context or for purposes violating patient confidentiality.

42. Any other comments?

The dataset is well-suited for machine learning and statistical analysis to identify predictors of MS progression.

Distribution

43. Will the dataset be distributed to third parties?

Yes, the dataset is publicly available.

44. How will the dataset be distributed (e.g., API, website)?

The dataset will be distributed through the Mendeley Data repository.

45. When will the dataset be distributed?

It is already publicly available.

46. Will the dataset be distributed under a license or terms of use?

Yes, under a Creative Commons Attribution 4.0 license.

47. Have any third parties imposed restrictions on the data?

Not mentioned.

48. Do any export controls or regulatory restrictions apply?

Not mentioned.

49. Any other comments?

Open availability and licensing facilitate reuse for research purposes.

Maintenance

50. Who will support/host/maintain the dataset?

The dataset is hosted on Mendeley Data.

51. How can the owner/curator of the dataset be contacted?

Contact information is not explicitly provided, but can likely be obtained via Mendeley Data.

52. Is there an erratum?

Not mentioned.

53. Will the dataset be updated?

Updates are not mentioned.

54. If the dataset relates to people, are there applicable data retention limits?

Not mentioned.

55. Will older versions of the dataset continue to be supported?

No older version on the dataset mentioned.

56. Is there a mechanism for others to extend/augment the dataset?

Not mentioned.

57. Any other comments?

The dataset appears static, with no explicit plans for updates or extensions.