**The Brazilian Epidemiologic Study of Twins and Families**

**Introduction**

Brazil is the most populous country in Latin America, with a current population of more than 200 million inhabitants. It has a universal health system (Sistema Único de Saúde – SUS) which was created in 1990. DATASUS was later established to manage routinely collected data from birth (SINASC) and death records (SIM) countrywide. Since then, other databases were incorporated into DATASUS, including data on hospitalisations, nutritional vigilance and other. Data originated from DATASUS is publicly available and can be downloaded by anyone who wishes to explore and analyse them.

Although the collection and management of population data through DATASUS has been consistent, the system has not yet integrated such databases, which constitute a challenge to more comprehensive epidemiologic studies. The most successful attempt to overcome this barrier was that of CIDACS, a centre dedicated to link databases in DATASUS to conduct epidemiologic studies. Although the data linkage process has been largely successful, generating a cohort of 114 million Brazilians (citation), CIDACS has not yet generated studies in the field of genetic epidemiology, where clusters of family members are studied to allow for inference on both and genetic and environmental aspects of human traits and conditions.

Twin studies have been widely successful in investigating associations between risk factors and disease while controlling for familial confounding (citation). They have also provided researchers with tools to be on the forefront of inference on causal relationships between exposures and outcomes (citation).

The Brazilian Epidemiologic Study of Twins and Families aims to conduct data linkage on Brazilian population databases originated from DATASUS, and to conduct twin and family studies by matching twin and other family members through machine learning algorithms. Our study is a multidisciplinary approach to tackle big health problems that are relevant not only to Brazilians but to the worldwide population. In our first data analysis, we will investigate sex differences on child mortality with the use of a co-twin control design with male-female twin pairs and generate Cox proportional hazard (PH) estimates.

**Data linkage and identification of twin pairs**

In order to conduct the data linkage component of our study, we have conducted the following steps:

1. Extracted data from SINASC and further extracting the twin cohort based on variable “Type of Pregnancy” (see table 1.1 below).
2. Created a machine learning algorithm that identifies the twins without information on type of pregnancy, based on similarity (correlated variables) with another record.
3. Created a machine learning algorithm to match twins in pairs based on variables described in Table 1.1.
4. Linked the SINASC and SIM databases using probabilistic linkage methods or when the birth registry number was recorded in the SIM database (missing data problem).
5. Generated a dataset with the twin cohort ready for epidemiological analysis, with an additional variable (family identifier)

Since DATASUS only makes available de-identified data, information on family relationships (such as a twin pair) is not readily available. To overcome that, we have developed an algorithm that predicts twins within a pair. This algorithm considers a number of maternal variables that are 100% correlated for twins in a pair. Table X provides an overview of the variables that were used to identify twin pairs from the SINASC database. Data from SINASC is available since XXXX, while data from SIM is available since

For some records, there is data on the type of pregnancy (single, twin or multiple) which in itself would identify the twins in the datasets. However, when there is missing data on this variable, the machine learning algorithm does also predict who is a twin and who is not in the database.

Table 1.1 – Variables used in machine learning algorithm

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Birth place | 9 - Ignored / 1 - Hospital / 2 - Other health establishment / 3 - Home / 4 - Other |
| Birth Place code | Code of the birth place health establishment |
| Suburb code | Code of the suburb of birth |
| Mother’s suburb | Code of the mother’s suburb of residence |
| City code | Code of the city of birth |
| Mother’s city | Code of the mother’s city of residence |
| Mother’s age | Mother’s age in years |
| Mother’s marital status | 1 - Single / 2 - Married / 3 - Widow / 4 - Divorced / 9 - Ignored |
| Mother’s education (years) | 1 - Zero / 2 - 1 to 3 years / 3 - 4 to 7 years / 4 - 8 to 11 years / 5 - 12 or more |
| Mother’s occupation | Mother’s occupation based on the Brazilian Occupation’s Classification |
| Mother’s alive children | Number of alive children of the mother |
| Mother’s deceased children | Number of deceased children of the mother |
| Mother’s gestation (weeks) | 9 – Ignored / 1 – Less than 22 weeks / 2 – 22 to 27 weeks / 28 to 31 weeks / 32 to 36 weeks / 37 to 41 weeks / 42 or more weeks |
| Type of gestation | 9 – Ignored / 1 – Single / 2 – Double (Twin) / 3 - Multiple |
| Prenatal consultations | Number of prenatal consultations |
| Birth date | DDMMYYY |

After conducting steps 1 to 5, the final result was a total of XX,XXX complete twin pairs. The accuracy of the algorithm to detect twin pairs was estimated at XX.XX%. (will need to explain the process here a bit better).

For those complete twin pairs with linked data from SINASC and SIM (birth and death records), the available variables can be found in table 1.2.

Table 1.2 – Other variables available from SINASC and SIM

|  |  |
| --- | --- |
| Variables on birth records (SINASC) database | |
| **Variable** | **Description** |
| Sex | 0 – Ignored / 1 – Male / 2 – Female |
| Apgar1 | Apgar at first minute of life |
| Apgar2 | Apgar at fifth minute of life |
| Birth weight | Weight at birth in grams |
| Variables on the death records (SIM) database | |
| Foetal death | 1 – Yes / 2 - No |
| Date of death | DDMMYYY |
| Marital status | 1 – Single / 2 – Married / 3 – Widow / 4 – Separated / 9 – Ignored |
| Education (years) | 1 – 0 years / 2 – 1 to 3 years / 3 – 4 to 7 years / 4 – 8 to 11 years / 5 – 12 or more years / 9 – Ignored |
| Perinatal death | 9 – Ignored / 1 – Before birth / 2 – During birth / 3 – After birth |
| Medical assistance | 9 – Ignored / 1 – Yes / 2 – Yes |
| Cause of death | Cause of death based on international classification of diseases |

**Future steps**

The consolidation of the linked dataset of Brazilian twin pairs through DATASUS will make possible the following subprojects, to be included in the NHMRC Ideas Grant:

1. ***Sex differences on child mortality: a population-based opposite-sex co-twin control study of Brazilian twins***

We will utilise the innovative opposite-sex co-twin control design, allowing for matching cases and controls for familial background and shared environmental factors. We will perform Cox Regression to establish proportional hazard estimates by analysing time-to-event related to mortality, using sex as the exposure variable. We will then investigate sex differences in the association between Apgar scores (both at first and fifth minute of life) and perinatal death as a categorical variable.

1. ***Further data linkage to allow more comprehensive epidemiologic studies***

We will link our dataset with other databases made available by DATASUS. Such databases include ISH (hospitalisations database), SISVAN (nutritional vigilance database) and other databases. We will work with CIDACS to access already available linked datasets with identifying information, resulting in the possibility of conducting wider family studies. Further data linkage will allow for more comprehensive and sophisticated twin study designs, including the utilisation of ICE FALCON (include reference), a novel twin study design aimed at improving inference on causal relationships.

Our algorithm to detect twin pairs will be rolled out to other international datasets that do not have available information on twin status.

1. ***Developing machine learning technology do classify/rank twin zygosity***

We will utilise within-pair analyses in variables available from step 2 in order to rank twin pairs on their probable zygosity status. This technology will allow for the utilisation of a variety of twin study designs, including co-twin control studies on same-sex twins, studies on genetic and environmental causes of trait variation and further utilisation of the ICE FALCON method.

**Impact**

Our project will have direct implications for clinical practice, and we plan to work closely with an established network of clinicians and researchers worldwide to ensure the relevance and impact of our research. The size of our study sample will potentially make this the largest opposite-sex twin study ever conducted in Brazil, if not the world. Our study design will allow for our findings to be generalizable to the whole studied populations, and not just twins.

Funding for our project will resource our multidisciplinary research teams in Brazil at UFBA (data linkage and machine learning), and Australia at the University of Melbourne (genetic epidemiology and data analysis) to conduct the steps outlined above.

Our research outcomes will include publications in major scientific journals, and we will make use of Twins Research Australia’s marketing and communications team to ensure the dissemination of our findings. We will have a knowledge translation strategy in place to allow for proper engagement with researchers, clinicians, parent organisations and policy-makers to discuss the implications of our findings in Brazil and worldwide.