Challenges in Health Data Management and Use

Scientific data management in Health and the Environment

Mauricio L. Barreto
Individual Based - Health Data Ecosystem

- Patient data
  - Health self-monitoring data
  - Wellness
  - Social networks...

- Administrative data
  - Risk management...

- Biological data
  - Omics data
    - Epigenomics
    - Genomics
    - Proteomics
    - Metabolomics
    - Lipidomics...

- Imaging data
  - MRI
  - DTI...

- Clinical data
  - Medical records

- Clinical trials

- Electronic Health records

- Patient registries

- Public Health records
  - Population risk management
  - Epidemiological studies

- BIOBANKS

Institution

National Academy of Sciences. 2011
Many challenges (General)

• Data heterogeneity (accuracy, format);
• Data fragmentation (multiple databases, multiple owners/stakeholders);
• Data availability (protection for commercial or cultural reasons, or related to personal privacy);
• Data conceptualisation (ontologies).
Many challenges (Health Data)

• Data handling (data management, data access, data quality, data querying, data sharing);
• Data privacy (Legal and ethical frames)
• Limitations on Data Linkage
• Limitations on Data Sharing
• Limitations on Data Access – Control mechanisms
Legal Frame

a) European Union - General Data Protection Regulation (GDPR)

b) Brazil – Law on Treatment and Protection of Personal Data (Law 13.709, From 14 August 2018)( Modified by Law 13.853, from 8 July 2019)
Considerations for safe data linkage environment

Data access approvals

Comprehensive approvals processes typically check that:

• There is a legal basis for data access
• There are appropriate security arrangements
• Data are used only for a specified purpose
• The requesting institution has appropriate credentials
• The ethics of the proposed study have been properly scrutinised
Considerations for safe data linkage environment

Researcher requirements

• Researchers have a responsibility to use data for *bona fide* purposes only
• Researchers should receive regular training in information governance
• Legal sanctions are in place where data are used inappropriately or without due care
Considerations for safe data linkage environment

Physical or virtual setting

Secure physical, or virtual, locations established for the processing and linkage of personal or potentially identifiable data, characterised by:

• Strict access arrangements
• Secure data transfer processes
• Restricted network and/or internet access
• Tight disclosure control procedures.
• Protected against outsider attacks or coercion
• Provide tangible reassurance on data security to the public
The Centre for Data and Knowledge Integration for Health (CIDACS): Linking Health and Social Data in Brazil

Barreto, ML\textsuperscript{1,2*}, Ichihara, MY\textsuperscript{1,2}, Almeida, BA\textsuperscript{1}, Barreto, ME\textsuperscript{1,3}, Cabral, L\textsuperscript{1}, Fiaccone, RL\textsuperscript{1,4}, Carreiro, RP\textsuperscript{1}, Teles, CAS\textsuperscript{1}, Pitta, R\textsuperscript{1}, Penna, GO\textsuperscript{1,5,6}, Barral-Netto, M\textsuperscript{2}, Ali, MS\textsuperscript{1,7,8}, Barbosa, G\textsuperscript{1}, Denaxas, S\textsuperscript{9}, Rodrigues, LC\textsuperscript{1,8}, and Smeeth, L\textsuperscript{1,8}

Abstract

The Centre for Data and Knowledge Integration for Health (CIDACS) was created in 2016 in Salvador, Bahia-Brazil with the objective of integrating data and knowledge aiming to answer scientific questions related to the health of the Brazilian population. This article details our experiences in the establishment and operations of CIDACS, as well as efforts made to obtain high-quality linked data while adhering to security, ethical use and privacy issues. Every effort has been made to conduct operations while implementing appropriate structures, procedures, processes and controls over the original and integrated databases in order to provide adequate datasets to answer relevant research questions. Looking forward, CIDACS is expected to be an important resource for researchers and policymakers interested in enhancing the evidence base pertaining to different aspects of health, in particular when investigating, from a nation-wide perspective, the role of social determinants of health and the effects of social and environmental policies on different health outcomes.
Data Platform

DATA ACCESS
- Dataset Request
- Dataset Information Search
- Dataset Visualization
- Data Analyses

DATA CURATION
- Data Entry and Exit
- Vocabulary and Metadata
- Anonimization
- Dataset Catalogue and Archiving

DATA PRODUCTION
- Ingestion
- Preparation and Processing
- Linkage
- Extraction

SECURITY

PRIVACY

ETHICS
The main data sources linked to build the 100 Million Cohort


- Family records, socio-demographic, schooling, housing, income and work CadUnico
- Records of the payments of the benefit of the income transfer program PBF
- Records of beneficiaries of the housing program MCMV
- Records of deaths SIM
- Records of notifications of 52 diseases SINAN
- Record monitoring of growth, development and food consumption SISVAN

Records of monitoring of compliance with PBF conditionalities (health and education) Condicionalidades
Records of beneficiaries of wells Wells
Records of live births, maternal and newborn characteristics SINASC
Records of Hospitalization AIH
COORTE DE NASCIMENTO - CIDACS

BASELINE

SINASC (birth)

CIDACS birth cohort

CADU (maternal information)

CADU (children information)

SINAN (compulsory notification disease)

SIH (hospitalization admission financed by the Brazilian National Health System)

SISVAN (nutritional conditions)

BOLSA FAMILIA PROGRAMME (governmental conditional cash transfer)

SIM (death)
Estrutura e componentes principais da coorte 100 milhões, fontes de dados e variáveis relevantes a serem vinculadas

<table>
<thead>
<tr>
<th>Components</th>
<th>Data source</th>
<th>Period</th>
<th>Number of Records</th>
<th>Relevant variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Cadastro Unico (CadUnico)</td>
<td>2001-2015</td>
<td>114,008,317</td>
<td>Social and economic; family composition; entry date.</td>
</tr>
<tr>
<td>Intervention (Exposure)</td>
<td>Cash Transfer Program-Bolsa Familia (BFP)</td>
<td>2004-2015</td>
<td>27,376,582</td>
<td>Start and end of data receipt of benefit, total value by family, age, and number of months received.</td>
</tr>
<tr>
<td>Conditionalities on health</td>
<td></td>
<td>2008-2015</td>
<td>255,737,815</td>
<td>Follow-up situation, frequency on the health services.</td>
</tr>
<tr>
<td>Conditionalities on education</td>
<td></td>
<td>2008-2015</td>
<td>154,032,061</td>
<td>Follow-up situation, frequency on school.</td>
</tr>
<tr>
<td>Housing Program (MCMV)</td>
<td></td>
<td>2009-2017</td>
<td>1,279,688</td>
<td>Marital status, date of receipt, monthly family income, municipality, date of birth, name of the enterprise, and subprogram.</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Deaths register in SIM (^1)</td>
<td>2000-2015</td>
<td>17,829,111</td>
<td>Type of death; date of death, date of birth, sex, race, education, duration of the pregnancy, single or multiple pregnancy, type of delivery, age of mother, gestational age, birth weight, and cause of death.</td>
</tr>
<tr>
<td></td>
<td>Births registered in SINASC (^2)</td>
<td>2001-2015</td>
<td>44,485,274</td>
<td>Mother’s data of birth, education, marital status, number of children and abortions, number of prenatal consultations, date of last menstruation, duration of gestation, type of pregnancy, type of delivery and child’s date of birth, sex, race, birth weight, presence of anomaly, who performed the childbirth among others.</td>
</tr>
<tr>
<td></td>
<td>Infectious diseases notified in SINAN (^3)</td>
<td>2001-2016</td>
<td>Tuberculosis (1,114,740), Leprosy (772,039), Zika (309,796), Chikungunya (344,523), Dengue (13,222,765)</td>
<td>Specific information of the infectious diseases notification.</td>
</tr>
<tr>
<td>Hospital admissions registered in SIH (^4)</td>
<td></td>
<td>2006-2017</td>
<td>136,598,439</td>
<td>Municipality of residence, medical specialty, hospital unit, type of care, race/color, reason for hospitalization, initial diagnosis, primary and secondary diagnosis (ICD), external causes.</td>
</tr>
</tbody>
</table>

1 Sistema de Informaca\(\) sobre Mortalidade.
2 Sistema de Informaca\(\) sobre Nascidos Vivos.
3 Sistema de Informaca\(\) de Agravos de Notificaca\(\).
4 Sistema de Informaca\(\)es Hospitale\(\).
<table>
<thead>
<tr>
<th>Author/year</th>
<th>Title</th>
<th>Journal</th>
<th>Exposure</th>
<th>Outcomes</th>
<th>Doi</th>
</tr>
</thead>
<tbody>
<tr>
<td>xx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pescarini JM et al. 2020</td>
<td>Conditional Cash Transfer Program and Leprosy Incidence: Analysis of 12.9 Million Families From the 100 Million Brazilian Cohort.</td>
<td>American Journal of Epidemiology</td>
<td>BFP receipt</td>
<td>Leprosy records (SINAN)</td>
<td>10.1093/aje/kwa127</td>
</tr>
<tr>
<td>Teixeira CSS et al 2020</td>
<td>Incidence of and Factors Associated With Leprosy Among Household Contacts of Patients With Leprosy in Brazil.</td>
<td>JAMA Dermatology</td>
<td>Baseline</td>
<td>Leprosy records (SINAN)</td>
<td>10.1001/jamadermatol.2020.0653</td>
</tr>
<tr>
<td>Pescarini JM et al 2020</td>
<td>Effect of a conditional cash transfer programme on leprosy treatment adherence and cure in patients from the nationwide 100 Million Brazilian Cohort: a quasi-experimental study.</td>
<td>Lancet Infectious Diseases</td>
<td>BFP receipt</td>
<td>Leprosy records (SINAN)</td>
<td>10.1016/S1473-3099(19)30624-3</td>
</tr>
<tr>
<td>de Andrade KVF et al 2019</td>
<td>Geographic and socio-economic factors associated with leprosy treatment default: An analysis from the 100 Million Brazilian Cohort.</td>
<td>PLoS Neglected Tropical Diseases</td>
<td>Baseline</td>
<td>Leprosy records (SINAN)</td>
<td>10.1371/journal.pntd.0007714 (Erratum in 10.1371/journal.pntd.0008723)</td>
</tr>
</tbody>
</table>
Obrigado!
Thank you!

mauricio.barreto@fiocruz.br
mauricio@ufba.br