Perguntas e Respostas (FAQ) sobre

Utilização de dados do Cidacs pelos projetos selecionados na chamada do Grand Challenges Explorations - Brasil: Ciência de dados para melhorar a saúde materno-infantil no Brasil

1. O que é a Coorte 100M SINASC-SIM?

A Coorte 100M SINASC-SIM CIDACS é uma base de dados para pesquisa em saúde, um dos produtos da Plataforma Coorte de 100 Milhões de Brasileiros.

2. Que dados estão disponíveis na Coorte 100M SINASC-SIM?

Os dados contêm informações sobre moradia, trabalho, escolaridade, renda, saneamento, deficiências, situação da população em situação de rua, indígenas, quilombolas, composição familiar, benefícios do Bolsa Família, nascimentos e óbitos de indivíduos, entre outros. Os dados se referem aos anos de 2006 a 2015.

3. Qual é a qualidade dos dados da Coorte 100M SINASC-SIM?

Todas as bases originais recebidas no CIDACS foram pré-processadas, incluindo limpeza, padronização e harmonização dos dados.

4. Como poderei acessar a Coorte 100M SINASC-SIM caso seja selecionado no GCE?

O candidato selecionado deverá preencher um formulário de requisição de pesquisa, com detalhes do pesquisador, do projeto, dos dados a serem usados e do parecer ético favorável. O pesquisador deverá aceitar as regras de privacidade e políticas de segurança de informação, e assinar um termo de responsabilidade de acesso aos dados. Após estes procedimentos, ele poderá acessar os dados localmente no CIDACS ou remotamente via VPN (Rede Particular Virtual/Virtual Private Network).

5. O que significa dados vinculados (Linkage)?

Dados vinculados são gerados a partir do processo de Record Linkage (vinculação de registros, em tradução livre), uma metodologia que calcula a similaridade de dados de forma determinística (quando há uma identificação única, como nos cadastros sociais que utilização o Número de Identificação Social) ou probabilística (por meio de informações variadas, como nome, data de nascimento e nome da mãe). Isso significa que uma nova base de dados foi criada a partir de duas ou mais bases de dados, após uma fase de pré-processamento, através da aplicação de um algoritmo de pareamento, em que informações sobre os mesmos indivíduos são integradas.

6. O que é anonimização de dados?

A anonimização é um procedimento para garantir a privacidade dos dados, por meio da aplicação de um algoritmo sobre variáveis (campos de informação) semi-identificadoras. Este processo assegura que indivíduos não possam ser identificados através de dados sensíveis, como nome ou endereço.

7. Os dados são individualizados ou agregados?

A Coorte 100M SINASC-SIM contém dados individualizados desidentificados e anonimizados.

8. Eu posso solicitar outros datasets (bancos de dados) ao CIDACS, se for selecionado?

A única base disponibilizada para o GCE é a Coorte 100M SINASC-SIM. Recortes dessa Coorte com menos anos, variáveis ou registros podem ser solicitadas pelos pesquisadores selecionados.

9. Eu posso levar a Coorte 100M SINASC-SIM para casa ou para o trabalho, se for selecionado?

De acordo com as normas de privacidade e segurança de dados do CIDACS, não é permitida a saída de dados do ambiente de análises do Centro. Deste modo, a Coorte 100M SINASC-SIM só poderá ser acessada localmente no CIDACS ou remotamente através de VPN.

10. Qual é o tamanho da Coorte 100M SINASC-SIM?

A versão mais atual da Coorte 100M SINASC-SIM possui aproximadamente 114 milhões de registros, 400 variáveis e ocupa 2 terabytes (armazenamento de disco).

11. Qual é a configuração mínima de máquina para se conseguir analisar a Coorte 100M SINASC-SIM?

A configuração irá depender do tipo de processamento. Aconselha-se sempre abrir e processar partes menores da base. Testes anteriores com bases semelhantes (número grande de registros e variáveis) indicam que análises descritivas na base inteira exigem pelo menos 256 GB de memória RAM. A maioria das ferramentas de análise de dados envolvendo pareamento (a exemplo do PSM – Propensity Score Matching) não possuem a capacidade de processamento paralelo e podem exigir pelo menos 1TB de memória RAM.

12. Que ferramentas estarão disponíveis para análise de dados?

O CIDACS oferece um ambiente de análises com máquinas virtuais com diversas configurações que podem ser alocadas de acordo com o plano de análise. As seguintes ferramentas estarão disponíveis: R, Python e STATA.

13. É possível usar R ou Python para analisar os dados da coorte?

Testes anteriores com bases semelhantes (número grande de registros e variáveis) indicam que a coorte é muito grande para a maioria dos algoritmos de distribuições básicas livres de R e Python. Sugerimos repartir a base em partes menores.

14. É possível usar STATA para analisar os dados da coorte?

O CIDACS vai disponibilizar o STATA versão 15, com capacidade de processamento paralelo (16, 24 ou 64 núcleos de CPU). Sugerimos repartir a base em partes menores.

15. Quais são os procedimentos de segurança de acesso aos dados?

Os pesquisadores selecionados devem se comprometer ao uso ético e seguro dos dados, que inclui utilizar os dados somente para a finalidade da pesquisa estabelecida, não distribuir os dados a terceiros, minimizar os riscos de acesso aos dados por pessoas não autorizadas. Deste modo, o selecionado deverá assinar os devidos termos de responsabilidade.

16. A Coorte 100M SINASC-SIM contém todos os dados ou variáveis das bases originais (fontes de dados)?

Não, a Coorte 100M SINASC-SIM contém apenas um subconjunto das variáveis originais, que são de interesse para pesquisa em saúde. No entanto, variáveis derivadas dessas bases foram adicionadas.

17. Posso receber uma amostra dos dados?

O CIDACS não poderá disponibilizar uma amostra de dados para candidatos não selecionados. Porém, o dicionário de dados, assim como outras informações podem ser encontrados nos links disponibilizados.

18. Como posso entrar em contato com o CIDACS para tirar dúvidas sobre a Coorte 100M SINASC-SIM?

Você pode entrar em contato com o CIDACS através do e-mail gcecidacs@fiocruz.org. Informações iniciais, incluindo este FAQ, estarão disponíveis em http://bit.ly/CidacsGCE