Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1DW8P

Title: Ecologia do sono e de fenótipos circadianos da população brasileira

Creator: Daniel Vartanian - **ORCID:** 0000-0001-7782-759X

Affiliation: Universidade de São Paulo (www5.usp.br)

Principal Investigator: Daniel Vartanian

Data Manager: Daniel Vartanian

Project Administrator: Daniel Vartanian

Contributor: Camilo Rodrigues Neto

Funder: Coordenação de Aperfeicoamento de Pessoal de Nível Superior (capes.gov.br)

Grant: 88887.703720/2022-00

Template: Digital Curation Centre (português)

Project abstract:

Teorias relacionadas ao sono e aos ritmos circadianos já estão bem consolidadas na ciência. No entanto, é necessário verificar e testar essas mesmas teorias em amostras mais abrangentes para obter um retrato mais preciso da expressão do sono e dos fenótipos temporais. Este projeto assume esse compromisso, tendo como objetivo mapear a expressão dos ciclos de sono-vigília e dos fenótipos circadianos da população adulta brasileira e investigar a hipótese de que a latitude está associada à regulação do ritmo circadiano. A hipótese da latitude se fundamenta na ideia de que regiões localizadas em latitudes próximas aos polos apresentam, em média, uma menor incidência de luz solar anual quando comparadas com regiões próximas da linha do equador (latitude 0°). Dessa forma, deduz-se que as regiões próximas ao equador apresentam um *zeitgeber* solar mais forte, o que, de acordo com as teorias da cronobiologia, deveria gerar uma maior propensão à sincronização dos ritmos circadianos dessas populações, reduzindo a amplitude e a diversidade dos fenótipos circadianos. Isso também daria a essas populações uma característica matutina quando comparadas com populações que vivem distantes da linha do equador. Para atingir os objetivos mencionados, o projeto irá contar com uma amostra de dados da expressão do ciclo sono-vigília da população brasileira composta por 120.265 respondentes que

abrange todos os estados brasileiros. Essa amostra de dados foi obtida no ano de 2017 e se baseia no Munich ChronoType Questionnaire (MCTQ), um questionário amplamente validado e utilizado para mensurar fenótipos circadianos a partir da expressão do ciclo sono-vigília de indivíduos em suas últimas quatro semanas. Com os resultados obtidos, espera-se contribuir com a validação de teorias da cronobiologia e gerar conhecimento sobre a regulação do ritmo circadiano e dos ciclos de sono-vigília na população brasileira.

Start date: 10-01-2023

End date: 07-31-2024

Last modified: 01-19-2024

Copyright information:

Ecologia do sono e de fenótipos circadianos da população brasileira

Este projeto parte de um banco de dados constituído no ano de 2017, **fora do âmbito de pesquisa**.

Os dados da amostra são primários e foram coletados por meio de um formulário público e online elaborado pelo principal autor do projeto (Daniel Vartanian). Este formulário permaneceu online entre os dias 20/04/2017 e 30/07/2019. Nele, havia uma versão padrão (*standard*) do *Munich ChronoType Questionnaire* (MCTQ) (Roenneberg *et al.*, 2003; Roenneberg *et al.*, 2012), uma escala amplamente validada e utilizada em estudos nas áreas do sono e da cronobiologia, cuja função é a mensuração de cronotipos (fenótipos circadianos) a partir da expressão nas últimas 4 semanas do ciclo sono-vigília de indivíduos. Um *snapshot d*o formulário de coleta, registrado no dia 18/10/2017 pela organização sem fins lucrativos Internet Archive, pode ser visualizado no endereço <https://web.archive.org/web/20171018043514/each.usp.br/gipso/mctq>.

No final do ano de 2017, por conta de uma entrevista realizada no programa intitulado Fantástico, da emissora de TV brasileira Globo, o formulário mencionado foi veiculado em rede nacional, resultando em um total de 118.068 respostas em uma janela de sete dias (de 15 a 21/10/2017) abrangendo todas as unidades federativas do Brasil. No total, 120.265 respostas foram coletadas no período em que o formulário permaneceu online. Devido ao formulário ainda ser um protótipo e não estar ligado a nenhuma pesquisa, os dados acabaram por ser coletados sem a presença de uma forte validação nos campos de resposta, e, por consequência, sua análise exige um trabalho árduo de limpeza e validação, motivo pelo qual eles ainda não foram analisados.

Como se pode notar, a amostra se deu por conveniência (método não probabilístico) e tem os indivíduos tanto como unidade de observação como unidade de análise.

As 120.265 entradas do conjunto de dados estão divididas em 94 variáveis. Essas variáveis podem ser agrupadas em cinco categorias:

- Dados de controle (*e.g.*, id; timestamp);
- Dados pessoais (*e.g.*, nome; sexo; CEP);
- Dados antropométricos/relacionados à saúde (*e.g.*, peso; altura; uso regular de medicamentos);
- Dados relacionados a hábitos (*e.g.*, se trabalha/estuda; períodos em que trabalha/estuda);
- Dados relacionados ao ciclo sono-vigília provindos do MCTQ (*e.g.*, hora local em que o indivíduo vai para cama em dias livres/de trabalho).

Um dos objetivos deste projeto está em fazer a limpeza e validação desses dados **em conjunto a anonimização dos respondentes**. A partir disso, eles serão utilizados para testar uma hipótese intitulada de *hipótese da latitude* ou *hipótese ambiental* (saiba mais em Leocadio-Miguel *et al.*, 2017) e poderão também servir a futuros projetos de pesquisa devidamente autorizados por um Comitê de Ética em Pesquisa (CEP) ligado à Comissão Nacional de Ética em Pesquisa (CONEP).

Serão também utilizados dados secundários de geolocalização pelo API (*Application Programming Interface*) Google Geocoding, dados solarimétricos do Atlas Brasileiro de Energia Solar (LABREN/ INPE) e dados do último censo demográfico disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Esses dados são necessários para análise e seu uso é explicado em outras seções deste documento.

No decorrer da análise, serão gerados outros conjuntos dados relacionados às variáveis computáveis do MCTQ (*e.g.*, duração do sono; tempo local do meio-sono) e a estatísticas pertinentes a modelos de regressão e a testes estatísticos variados. Tabelas de frequência, estatísticas descritivas e diferentes tipos de saídas gráficas, também serão geradas de forma a representar a amostra e os resultados obtidos pelo estudo.

Conforme colocado anteriormente, este projeto parte de um banco de dados constituído no ano de 2017, **fora do âmbito de pesquisa**. A descrição abaixo se refere aos dados criados pelo processamento e análise dos dados brutos e aos dados secundários que serão utilizados.

O processamento dos dados brutos do banco de dados será realizado com base no programa para ciência de dados delineado por Hadley Wickham & Garrett Grolemund. Os dados serão manipulados utilizando a linguagem de programação R e Python em conjunto com o IDE (*Integrated Development Environment*) RStudio.

O processamento das variáveis computáveis do MCTQ será realizado por meio do pacote R { mctq}, pacote gratuito, de código aberto, revisado por pares pela iniciativa rOpenSci e disponível na The Comprehensive R Archive Network (CRAN).

Após a limpeza e validação dos dados, serão coletados e cruzados dados de geolocalização pelo API (*Application Programming Interface*) Google Geocoding (Google, n.d.) de forma a obter as coordenadas de latitude e longitude próximas das residências dos respondentes. Isso será realizado com base nos Códigos de Endereçamento Postal (CEP) fornecidos. Após a obtenção da geolocalização, serão então coletados e cruzados dados solarimétricos do Atlas Brasileiro de Energia Solar (Pereira *et al.*, 2017) produzidos pelo Laboratório de Modelagem e Estudos de Recursos Renováveis de Energia (LABREN) do Instituto Nacional de Pesquisas Espaciais (INPE) de forma a obter a irradiação solar aproximada do local de residência do respondente no momento de sua resposta. Esses dados são importantes para a modelagem do fenômeno de e*ntrainment* relacionado à hipótese a ser testada.

Os dados válidos ainda terão que ser balanceados a partir da geração de subamostras aleatórias ajustadas às proporções das regiões brasileiras. Isso será feito com base nos dados do último censo demográfico disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Essas subamostras serão geradas conforme a granularidade requisitada por cada análise (*e.g.*, subamostra de representatividade nacional; subamostra do estado de São Paulo).

Os dados finais serão **anonimizados**, divididos e armazenados conforme as seguintes entidades/conjuntos de dados:

- *subject*: entidade na qual serão armazenadas variáveis relacionados a dados pessoais dos respondentes (*e.g.*, sexo; idade);
- *health*: entidade na qual serão armazenadas variáveis relacionadas à saúde dos respondentes (*e.g.*, altura; peso; uso diário de medicamentos);
- *habits*: entidade na qual serão armazenadas variáveis relacionadas aos hábitos de estudo e trabalho dos respondentes. (*e.g.*, se trabalha/estuda; períodos em que trabalha/estuda);
- *mctq*: entidade na qual serão armazenadas variáveis relacionadas à escala de cronotipo utilizada (MCTQ) (*e.g.*, hora local em que o indivíduo vai para cama em dias livres/de trabalho);
- *geolocation*: entidade na qual serão armazenadas variáveis relacionadas à geolocalização dos respondentes (*e.g.*, país; CEP; latitude central do CEP; longitude central do CEP);
- *solarimetry*: entidade na qual serão armazenadas variáveis relacionadas à irradiação solar média conforme a latitude e longitude centrais do CEP dos respondentes (*e.g.*, irradiação global horizontal).

Os dados finais serão organizados conforme um *Entity Relationship Diagram* (ERD) elaborado na ferramenta dbdiagram.io. Esse modelo pode ser visualizado no link: <https://dbdiagram.io/d/brchrono-5e6c84484495b02c3b883aff>. É importante notar que, apesar desses dados estarem em um esquema relacional de banco de dados, a normalização desse esquema não foi considerada, dado que se trata de conjuntos de dados estáticos e de baixa complexidade.

Todos os procedimentos de limpeza, validação, transformação e análise dos dados ficarão contidos em um

repositório hospedado na plataforma GitHub que representará o compêndio da pesquisa. Esse repositório será estruturado de acordo com o padrão de pacotes da linguagem de programação R. Essa forma de organização foi inspirada a partir de um artigo publicado por Ben Marwick, Carl Boettiger e Lincoln Mullen (2018).

Os procedimentos mencionados acima estarão ligados a *notebooks* computacionais baseados no sistema de publicação Quarto. Isso irá garantir uma melhor organização e reprodutibilidade não só das análises, mas também dos artigos que irão acompanhá-las.

A documentação dos dados será realizada por meio de um pacote R criado com o propósito de facilitar o compartilhamento deles. Esse pacote será acompanhado de um website documental elaborado utilizando os pacotes da linguagem de programação R {roxygen2} e {pkgdown}. Nele o usuário irá encontrar artigos e orientações gerais sobre os dados, a documentação de todas as funções/algoritmos auxiliares e a documentação de todas as entidades/conjunto de dados relacionados ao ERD mencionado anteriormente.

Os dados também terão suas informações (metadados) divulgadas no repositório USP ao final do estudo de forma a atender aos princípios FAIR (*Findability*; *Accessibility*; *Interoperability*; *Reuse*). Dado o aspecto único do conjunto dados a ser produzido, pretende-se também publicar um artigo exclusivo sobre eles em uma revista voltada à cronobiologia.

Todas as questões éticas serão administradas conforme as normas e regulamentos do Conselho Nacional em Ética em Pesquisa (CONEP) do Conselho Nacional de Saúde (CNS) e da Lei de Proteção de Dados Pessoais (LPDP - Lei nº 13.709/2018). Desta forma, este estudo será submetido pela avaliação do Comitê de Ética em Pesquisa (CEP) da Escola de Artes, Ciências e Humanidades (EACH) da Universidade de São Paulo (USP) a fim de se obter a aprovação necessária para sua implementação.

É importante notar que o conjunto de dados utilizados neste estudo foi criado **fora do âmbito de pesquisa** e que os dados finais serão todos **anonimizados**. Além disso, deve-se notar também que, ao preencherem o formulário de coleta, os respondentes foram informados que seus dados seriam guardados em sigilo e seriam somente utilizados para fins de pesquisa, havendo, portanto, um registro a respeito de seu uso futuro. Isso pode ser verificado no *snapshot* do formulário de coleta feito pela organização sem fins lucrativos Internet Archive no dia 18/10/2017, momento no qual cerca de 98,173% dos dados da amostra foram coletados. Ele pode ser visualizado em: <https://web.archive.org/web/20171018043514/each.usp.br/gipso/mctq>.

A partir das normas do CONEP, entende-se que este estudo é admissível em razão de sua relevância social, singularidade e da presença de apenas um risco mínimo de violações e vazamentos de dados de participantes por terceiros. Esse risco se justifica pelo benefício esperado (Item V.1.A da Resolução nº 466, de 12 de dezembro de 2012 do CNS). Entende-se também que este estudo pode ser enquadrado como uma pesquisa de interesse estratégico para o SUS (Resolução nº 580, de 22 de março de 2018), visto que os dados finais representarão o maior e mais atualizado mapa sobre o ciclo sono-vigília da população brasileira até o momento. Ainda, é importante notar que este estudo já está de acordo com os termos da consulta pública para a propositura de resolução sobre o uso de bancos de dados com finalidade de pesquisa científica envolvendo seres humanos (Saiba mais em <https://redcap.saude.gov.br/surveys/?s=F9CN8JYXKD>).

Caso ocorra alguma questão ética durante o estudo, essas serão trazidas ao CEP da EACH-USP para orientação e, se necessário, administração do caso.

Todos os materiais produzidos por este estudo, com exceção dos dados dos respondentes, serão acompanhados de uma licença aberta. O código-fonte será distribuído por meio da licença de código aberto MIT. Documentos e outros materiais do tipo serão disponibilizados por meio da licença Atribuição 4.0 Internacional da Creative

Commons (CC BY 4.0).

Os dados dos respondentes são de propriedade deles mesmos. Logo, eles só serão compartilhados conforme autorização de estudo por parte do Conselho Nacional em Ética em Pesquisa (CONEP) e de seus controladores. As outras seções deste documento versam mais a respeito disso.

Todos os dados utilizados e produzidos neste estudo serão armazenados nos formatos CSV (*Comma-Separated Values*) e/ou RDA (*R Data File*) em um *bucket* na ferramenta Google Cloud Storage. Esse armazenamento terá somente acesso privado e irá contar com uma camada de criptografia fornecida pelo próprio serviço de armazenamento em nuvem. Entretanto, para tornar ainda mais seguro os dados, eles serão armazenados na nuvem com uma camada adicional de criptografia realizada localmente por meio da geração de um par de chaves RSA (Rivest-Shamir-Adleman) exclusiva do projeto. Isso evitará que o próprio prestador de serviço de armazenamento consiga acessar o conteúdo deles.

É importante notar que a tecnologia de armazenamento em nuvem do Google Cloud Storage apresenta redundância e está menos sujeita a defasagens tecnológicas, ao contrário das mídias físicas. Todavia, por segurança, uma cópia criptografada dos dados, será mantida localmente com os controladores.

Os dados utilizados e produzidos por este projeto de pesquisa irão contar com duas camadas de segurança. A primeira está relacionada com a necessidade de autorização do controlador para acesso de terceiros aos dados armazenados em nuvem, além do envio de uma chave de criptografia gerada pelo próprio serviço de armazenamento (Google Cloud Storage). A segunda está relacionada com a necessidade de uma chave RSA (Rivest-Shamir-Adleman) adicional para abrir os arquivos de dados que estão armazenados na nuvem.

Logo, para se possa obter acesso aos dados, será necessário a autorização do usuário pelo controlador dos dados no serviço de armazenamento em nuvem e o envio e instalação de diversas chaves de criptografia. No caso de compartilhamento de dados para pesquisas futuras, esse acesso somente será permitido após o atendimento de todos os requisitos, conforme colocado em outras seções deste documento.

Como exceção, o acesso temporário aos dados poderá ser permitido, por meio de um acordo de responsabilidade, a revisores de artigos de revistas científicas, com a finalidade de reproduzir e verificar os resultados obtidos pelo estudo.

Por se tratar de uma amostra única, é difícil prever todos os seus possíveis usos. Logo, a princípio, todos os dados brutos e processados serão mantidos e preservados por um prazo indeterminado.

O compartilhamento dos dados também estará disponível por um prazo indeterminado, desde que todos os requisitos para sua reutilização sejam atendidos. Veja as outras seções deste documento para saber mais sobre os requisitos.

Os dados serão preservados por prazo indeterminado por meio do serviço de armazenamento em nuvem Google Cloud Storage e por meio de uma cópia criptografada dos dados mantida localmente com o controlador.

As chaves RSA do projeto estarão seguras tanto localmente como na nuvem, por meio do gerenciador de senhas 1password.

Há tanto benefícios como riscos no armazenamento dos dados em tecnologias de nuvem. Entre os benefícios está que essas tecnologias apresentam redundância e estão menos sujeitas a defasagens tecnológicas, ao contrário das mídias físicas. Entre os riscos estão possíveis violações e vazamentos de dados. Esses riscos, porém, são mínimos,

e os dados serão sempre armazenados na nuvem com pelo menos duas camadas de criptografia.

O compartilhamento dos dados, quando aprovados, será realizado por meio de uma autorização de acesso personalizada ao *bucket* do projeto no serviço de nuvem Google Cloud Storage e pelo envio de chaves de criptografia personalizadas aos solicitantes. Esse conjunto de chaves será necessário para que os dados possam ser acessados e destrancados. Essas chaves poderão ser facilmente canceladas a qualquer momento pelos controladores, caso eles achem necessário.

Os solicitantes só poderão acessar os dados de forma programática e não poderão armazená-los localmente. Exceção feita, é claro, a armazenamentos pontuais ocorridos durante o processo das análises e para visualização das tabelas. Isso exigirá dos solicitantes um conhecimento em programação. Os controladores oferecerão todo o auxílio necessário para a instalação das chaves e acesso ao serviço de armazenamento em nuvem.

Sim.

O compartilhamento dos dados só será realizado por meio de uma solicitação e aprovação por parte dos controladores, se todos os aspectos éticos previstos na legislação brasileira para pesquisas envolvendo seres humanos sejam cumpridos. Ou seja, será necessário que a solicitação para o compartilhamento dos mesmos inclua um estudo que possa ser submetido à aprovação pelo Conselho Nacional em Ética em Pesquisa (CONEP).

Se os controladores aprovarem o uso dos dados, esses somente serão liberados após a confirmação de aprovação de um Comitê de Ética em Pesquisa (CEP) ligado ao CONEP. Essa confirmação será realizada via consulta pela internet na Plataforma Brasil feita por meio do Certificado de Apresentação para Apreciação Ética (CAAE) e pelo número do parecer do CEP. O controle e gerenciamento dos dados é tratado em outra seção deste documento.

Segundo as normas do Conselho Nacional em Ética em Pesquisa (CONEP), recai sob o orientador do estudo e pesquisador principal, Camilo Rodrigues Neto, a responsabilidade pela implementação deste plano de dados.

O controle dos dados finais ficará sob a responsabilidade dos investigadores principais e gerenciadores de dados descritos neste plano.

Este plano conta com recursos fornecidos pela Universidade de São Paulo (USP) e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Entre esses recursos está uma bolsa de mestrado já concedida ao autor do plano de dados (Daniel Vartanian) (Número da concessão: 88887.703720/2022-00).

Além da bolsa, há poucos custos para a implementação das ações programadas neste documento. Será necessário somente o uso de um computador de configurações medianas e da assinatura do serviço de armazenamento Google Cloud Storage. Esses serão financiados pela equipe de pesquisa.
