Title: The Brazilian Epidemiologic Study of Twin and Multiple Families (BEST)

Section I. What is your idea?

A twin or multiple birth is a major risk factor for prematurity, low birthweight and perinatal mortality. Twins and other multiples are much more likely to be born pre-term than singletons (64% vs. 7%)[1]. The proportion of low birthweight babies is substantially higher for twins (54%) and other multiples (98%) than for singletons (5%), and a perinatal death is 4 and 8 times more likely for twins and multiples compared to singletons.[1]

However, little is known about how risk factors for prematurity and low birthweight and their effects on mortality differ between singletons, twins and multiples. The inception of a Brazilian population-based cohort of twins, multiples and their families as a high-risk group and its comparison with singletons is essential to address this research gap and improve the health care of mothers and infants in Brazil and worldwide. It can also provide valuable insights to help formulating evidence-based guidelines and policies that will also have a substantial positive impact to the health care system of Brazil, where nearly 60,000 twins and multiples are born each year.

We propose to create a Brazilian Epidemiologic Study of Twin and Multiple Families (BEST), a cohort study of twins, multiples and their mothers based on linked population databases available from CIDACS. We will develop predictive models for preterm birth, low birthweight, mortality and causes of death in twin, multiple and singleton births. We will also make use of the uniqueness of a powerful study design involving male-female twin pairs to investigate sex differences in such poor outcomes while controlling for familial (genetic and environmental) factors as confounders.

This comprehensive epidemiological approach will yield findings that are relevant, not only for the health care of twins, but to the whole population. The main research questions to be answered with the study are:

1) What are the increased risks associated with preterm and low birthweight births for twins and other multiples compared with singletons? How does the prevalence of preterm and low birthweight births in twins and multiples compare with singletons?
2) Are risk factors for preterm and low birthweight births similar for singletons, twins and other multiples?
3) Is there an association between birth plurality (single, twin or multiple birth) and mortality? Are prematurity and low birthweight competing risks for mortality for each of these groups? Do main causes of death differ between twins, multiples and singletons?
4) Is there an association between number of prenatal consultations and birth plurality?
5) Are sex, prematurity, low birthweight associated with mortality after matching for age and controlling for familial factors and other confounders?

This will be an innovative and collaborative approach to investigating twin and other multiple births as major risk factors for important health-related complications from conception through to early-life and childhood. We aim to design high accurate algorithms to ascertain twin pairs in the provided linked datasets while guaranteeing de-identification of individuals. This will enable us to use an innovative twin study design to investigate sex differences in poor perinatal health outcomes in a way not otherwise possible.

Our study will give us valuable insights into the Brazilian population and allow us to compare our results with those from ongoing similar studies conducted by our international partners through a global consortium of researchers, clinicians and parent organisations aimed at recommending evidence-based prenatal care guidelines specific for twin and multiple births to improve their health care. This will also be informed by our analysis on how singleton, twin and multiple births differ in the number of prenatal consultations.

We will bring a multi-disciplinary team and collaborators including epidemiologists, biostatisticians, clinicians and data scientists to achieve our goals and potentialize future opportunities in the analysis of Brazilian population data. We will be supported by a Centre of Excellence in Twin Research in Australia through our ongoing collaborations. The project's team will include experts in machine learning and data linkage, and leaders of twin research in Brazil and collaborators in other countries.

---

[1] Australian Institute of Health and Welfare. *Australia's mothers and babies 2015 — in brief.* Canberra, Australia. 2015.

Section II. How will you test it?

Initially, we will require from CIDACS a linked dataset (with the SINASC and SIM databases) with all subjects born between 2006 and 2015, including all available variables from those databases. Once we are granted access to the required CIDACS linked datasets, we will perform the following steps: 1) Technology development (6 months). 2) Data analysis - see detailed plan below (6 months). 3) Production of manuscripts, dissemination of findings and recommendations (4 months). 4) Planning of new projects and collaborations under phase 2 (2 months).

In ***step 1***, we will apply machine-learning techniques to develop statistical models using conditional regression techniques to probabilistically link de-identified twins in the same pair. We will then use data simulation procedures to assess the accuracy of our models and select the model which best predict twin pairs. These models will use variables that are perfectly correlated for twin pairs to improve accuracy, such as: birth plurality, type of birth delivery, date of birth, maternal age, gestational age, number of prenatal consultations, number of pregnancies (mother), mothers' age and potentially other variables.

In ***step 2***, we will conduct data analyses to answer each research question in Section 1, as follows: **Research question 1**: To estimate risks of outcomes (preterm and low birthweight births,) for twins and other multiples compared with singletons using robust statistical models (Poisson models fitted using generalized estimating equations and accounting for clustering between siblings). **Research question 2:** To evaluate the effect of risk factors (maternal age, paternal age, maternal parity, gestational age and Apgar scores) individually and as mediating factors on the risks estimated in Research Question 1. **Research question 3:** To perform a survival analysis to assess mortality risk using birth plurality (singleton, twin or multiple) as the exposure. Then, fit a competing risk model including low birthweight and preterm as competing events to investigate their effect on mortality. **Research question 4:** To quantify the association of the number of prenatal consultations with type of birth using logistic regression modelling. **Research question 5:** To quantify the association of sex with prematurity and low birthweight using conditional logistic regression modelling, while controlling for familial factors.

In ***step 3***, we will publish our findings (including comparisons with ongoing similar studies conducted by our international collaborators) in highly ranked journals and disseminate a white paper with evidence-based recommendations on how the findings from our study can be used to inform health care policy and practice in Brazil and internationally. We will present our findings and recommendations in international conferences and congresses. Master dissertations and PhD theses will be also produced as a result from our study.

Our initiative will make future national and international collaborative projects possible with additional funding, and those will be discussed and planned with existing and new partners in ***step 4***. First, the generated data can be linked with other Brazilian population databases from DATASUS (such as SISVAN and SIH), improving the accuracy of our predictive models and the monitoring of twins and multiples longitudinally, including integration with Brazilian health care monitoring units and the recently established Brazilian Twin Registry. Second, our algorithm to predict male-female twin pairs in de-identified population datasets to efficiently investigate the role of sex differences in a number of human traits and conditions can be rolled out to the additional linked datasets, generating a variety of new twin studies which control for familial factors. Third, the available data on Brazilian twins will permit the development of machine learning tools for the classification of zygosity of de-identified twins using sophisticated statistical modelling. This technology will be tremendously valuable to the application of novel causal inference models using twin data[2] available in Brazil and internationally, providing a further step towards the next generation of genetic epidemiologic studies.

The total requested amount is US$99,930. It includes US$49,890 for personnel costs and subcontracts, including one PI with expertise in maternal and child health, and two Co-PIs with expertise in epidemiology/biostatistics and data science (US$8,960 each), one Post-doc student ($7,616) and one PhD student (U$6,854) for data linkage and analysis, external consultants such as clinicians and biostatisticians ($5,260), and a subcontract for development of a website with data visualisation features (US$3,680). Around US$12,980 will be used for purchase of equipment, $1,800 for supplies and consumables, and US$34,860 for travel costs. Equipment costs will include the purchase of desktop computers to be used in the data analyses and other equipment dedicated to technology development and the remote access to the CIDACS computational infrastructure. Travel costs will cover flight tickets and accommodation for an initial planning meeting at the beginning of activities with the project's team and collaborators, conference presentations, and meetings with existing and prospect national and international collaborators to plan for phase 2.

---

[2] Li S, Wong EM, Bui M, et al. Inference about causation between body mass index and DNA methylation in blood from a twin family study. *International Journal of Obesity - in press.* 2018.

Título: Estudo Epidemiológico Brasileiro de Famílias de Gêmeos e Múltiplos

Seção I. Qual é a sua ideia?

Um nascimento gemelar ou múltiplo é um importante fator de risco para prematuridade, baixo peso ao nascer e mortalidade perinatal. Gêmeos e outros múltiplos têm muito mais chances de nascer pré-termo do que os únicos (64% a 7%).[1] A proporção de bebês com baixo peso ao nascer é substancialmente maior para gêmeos (54%) e outros múltiplos (98%) do que para únicos (5%), e uma morte perinatal é 4 e 8 mais provável para gêmeos e múltiplos, respectivamente, do que para únicos[1].

No entanto, sabe-se pouco sobre como os fatores de risco para prematuridade e baixo peso ao nascer e seus efeitos sobre a mortalidade diferem entre únicos, gêmeos e múltiplos. A criação de uma coorte de base populacional de gêmeos, múltiplos e suas famílias como um grupo de alto risco e sua comparação com filhos únicos é essencial para preencher essa lacuna e melhorar os cuidados de saúde materno-infantil no Brasil e no mundo. Também fornecerá informações valiosas para ajudar a formular diretrizes e políticas baseadas em evidências que também terão um impacto substancial no Sistema Único de Saúde do Brasil, onde perto de 60.000 gêmeos e múltiplos nascem a cada ano.

Propomos a criação do Estudo Epidemiológico Brasileiro de Famílias de Gêmeos e Múltiplos um estudo de coorte de gêmeos, múltiplos e suas mães baseado em bancos de dados populacionais disponibilizados pelo CIDACS. Nós iremos desenvolver modelos preditivos para parto prematuro, baixo peso ao nascer, mortalidade e causas de morte entre nascimentos de gêmeos, múltiplos e únicos. Também faremos uso da singularidade e robustez de um design de estudo envolvendo pares de gêmeos masculino-feminino para investigar as diferenças entre os sexos em desfechos desfavoráveis controlando fatores familiares (genéticos e ambientais) como fatores de confusão.

Essa abordagem epidemiológica abrangente produzirá evidências relevantes, não apenas para a saúde dos gêmeos, mas para toda a população. As principais perguntas a serem respondidas pelo estudo são:

1) Quais são os riscos aumentados associados a partos prematuros e com baixo peso ao nascer para gêmeos e outros múltiplos em comparação com os filhos únicos? Como a prevalência de nascimentos prematuros e de baixo peso ao nascer em gêmeos e múltiplos se compara a filhos únicos?
2) Os fatores de risco para partos prematuros e com baixo peso ao nascer são semelhantes para os filhos únicos, gêmeos e outros múltiplos?
3) Existe associação entre o tipo de gravidez (única, gemelar ou múltipla) e mortalidade? A prematuridade e o baixo peso ao nascer competem com os riscos de mortalidade para cada um desses grupos? As principais causes de morte diferem entre gêmeos, múltiplos e únicos?
4) Existe associação entre o número de consultas pré-natal e o tipo de gravidez?
5) Sexo, prematuridade e baixo peso ao nascer estão associados à mortalidade após pareamento por idade e controle de fatores familiares e outros fatores de confusão?

Esta será uma abordagem inovadora e colaborativa que possibilitará a investigação do nascimento de gêmeos e outros múltiplos como os principais fatores de risco para importantes complicações relacionadas à saúde, desde a concepção até a infância. Desenvolveremos um algoritmo de última geração para agrupar pares de gêmeos nos conjuntos de dados vinculados fornecidos garantindo a não identificação dos indivíduos, permitindo a utilização de um inovador design de estudo com pares de gêmeos, visando investigar diferenças entre sexos em desfechos perinatais desfavoráveis. Nosso estudo trará informações valiosas sobre a população brasileira e nos permitirá comparar nossos resultados com os de estudos semelhantes em andamento conduzidos por nossos parceiros internacionais em um consórcio global entre pesquisadores, profissionais da área da saúde clínica e organizações de pais, que visa melhorar a atenção à saúde de gêmeos, outros múltiplos e suas mães, a fim de sugerir a implementação de diretrizes de cuidados de saúde específicas para nascimentos gêmeos e múltiplos. Nossas recomendações também serão viabilizadas pela análise de como os nascimentos únicos, gêmeos e múltiplos diferem em relação ao número de consultas de pré-natais.

A equipe de projeto e de colaboradores terá um caráter multidisciplinar, incluindo epidemiologistas, bioestatísticos, obstetras e cientistas de dados para alcançar nossos objetivos e potencializar futuras oportunidades na análise de dados populacionais brasileiros. Seremos apoiados por um Centro de Excelência em Pesquisa de Gêmeos na Austrália através de nossas colaborações em andamento. A equipe do projeto inclui um especialista em aprendizado automático e vinculação de dados, além de líderes de pesquisa de gêmeos no Brasil e colaboradores em outros países.

---

[1] Australian Institute of Health and Welfare. *Australia's mothers and babies 2015 — in brief.* Canberra, Australia. 2015.

Seção II – Como irá testá-la?

Inicialmente, solicitaremos ao CIDACS um conjunto de dados vinculados (com as bases de dados do SINASC e SIM) com todos os participantes nascidos entre 2006 e 2015, incluindo todas as variáveis disponíveis nessas bases de dados. Com acesso aos conjuntos de dados vinculados do CIDACS, executaremos as seguintes etapas: 1) Desenvolvimento tecnológico (6 meses); 2) Análise de dados - plano detalhado abaixo (6 meses); 3) Produção de manuscritos, divulgação de resultados e recomendações (4 meses) e; 4) Planejamento de novos projetos e colaborações da fase 2 (2 meses).

Na etapa 1, aplicaremos técnicas de aprendizado automático para desenvolver modelos estatísticos usando regressão condicional para agrupar probabilisticamente gêmeos no mesmo par de forma não identificável. Em seguida, usaremos procedimentos de simulação de dados para avaliar a precisão de nossos modelos e selecionaremos o modelo que melhor prever os pares de gêmeos. Os modelos utilizarão variáveis perfeitamente correlacionadas para pares de gêmeos para melhorar a precisão, tais como: pluralidade de nascimentos, tipo de parto, data de nascimento, idade materna, idade gestacional, número de consultas de pré-natal, número de gestações, idade materna e potencialmente outras variáveis.

Na **etapa 2**, a análise de dados irá responder a cada pergunta da Seção 1, da seguinte maneira: **Pergunta 1)** Estimar os riscos de desfechos (nascimentos prematuros e de baixo peso ao nascer) para gêmeos e outros múltiplos comparados com nascimentos únicos usando modelos estatísticos robustos (modelos de Poisson usando equações estimativas generalizadas e considerando agrupamento entre irmãos). **Pergunta 2:** Avaliar o efeito de fatores de risco (idade materna, idade paterna, paridade materna, idade gestacional e Apgar) individualmente e como mediadores dos riscos estimados na pergunta de pesquisa 1. **Pergunta 3:** Realizar análise de sobrevida para avaliar o risco de mortalidade usando pluralidade de nascimento (individual, gêmeo ou múltiplo) como a exposição. Desenvolver um modelo de risco competitivo, incluindo baixo peso ao nascer e prematuridade, como eventos concorrentes, para investigar seus efeitos sobre a mortalidade. **Pergunta 4:** Quantificar a associação do número de consultas de pré-natal com o tipo de gravidez utilizando a modelo de regressão logística. **Pergunta 5:** Verificar a associação entre sexo e prematuridade e baixo peso ao nascer usando modelo de regressão logística condicional, controlando os fatores familiares.

Na etapa 3, publicaremos nossos resultados (incluindo comparações com estudos similares em andamento conduzidos por nossos colaboradores internacionais) em revistas de alto impacto e divulgaremos relatório de recomendações baseadas em evidências sobre como as descobertas do estudo podem ser usadas para melhorar políticas de saúde e práticas no Brasil e no mundo. Apresentaremos nossas descobertas e recomendações em conferências e congressos internacionais. Dissertações de mestrado e teses de doutorado também serão produzidas como resultado de nosso estudo.

O estudo tornará possível futuros projetos nacionais e internacionais de colaboração com financiamento adicional, e esses serão discutidos e planejados com novos e existentes parceiros na **etapa 4**. Primeiro, os dados gerados podem ser vinculados a outros bancos de dados do DATASUS (como SISVAN e SIH), permitindo a melhoria da precisão de nossos modelos preditivos para monitorar gêmeos e múltiplos longitudinalmente, incluindo a integração com as unidades brasileiras de monitoramento de saúde e o recém-criado Registro Brasileiro de Gêmeos. Segundo, nosso algoritmo para agrupar pares de gêmeos masculino-feminino em dados populacionais não identificáveis para investigar diferenças entre sexos em vários traços e condições humanas pode ser aplicado às novas bases vinculadas, gerando uma variedade de novos estudos com gêmeos que permitem o controle para fatores de risco familiares. Terceiro, os dados de gêmeos brasileiros permitirão o desenvolvimento de ferramentas de aprendizado automático para a classificação da zigosidade de gêmeos usando modelagem estatística sofisticada. Essa tecnologia será extremamente valiosa para a aplicação de novos modelos de inferência causal com dados populacionais de gêmeos[2] em databases disponíveis no Brasil e em outros países, fornecendo mais um passo em direção à próxima geração de estudos na área de epidemiologia genética.

O valor total solicitado é de US$ 99.930. O valor inclui US$49.890 para custos de pessoal e subcontratos, incluindo o PI especialista em saúde materno-infantil, e dois Co-IPs especialistas em epidemiologia/bioestatística e ciência de dados (US$8.960 cada), um estudante de pós-doutorado (US$7.616), um estudante de doutorado ($6.854) para vinculação e análise de dados, consultores externos (US$5.260) e o subcontrato para desenvolvimento de um website de visualização de dados ($3.680). Cerca de US$12.980 serão usados na compra de equipamentos, US$1.800 em suprimentos e consumíveis e US$34.860 para despesas de viagem. Os custos de equipamento incluirão a compra de computadores de mesa para análise de dados e outros equipamentos dedicados ao desenvolvimento de tecnologia e o acesso remoto à infraestrutura computacional do CIDACS. As despesas de viagem cobrirão passagens aéreas e acomodação para uma reunião de planejamento inicial no início das atividades com a equipe do projeto, apresentações em conferências e reuniões com os existentes e potenciais colaboradores nacionais e internacionais para planejar a fase 2.

---

[2] Li S, Wong EM, Bui M, et al. Inference about causation between body mass index and DNA methylation in blood from a twin family study. *International Journal of Obesity - in press.* 2018.