Analysis of birth cohort studies in BayesDB
Leonardo Casarsa, Belhal Karimi and Vikash K. Mansinghka

Abstract
The Bill & Melinda Gates Foundation has a multi-year project on knowledge integration for Healthy Birth, Growth, and Development (HBGDkI). The core aim is to make it possible to perform empirically grounded policy design, advocacy, and field treatment around growth insults such as stunting due to malnutrition. Tactics include gathering data on >5M children; funding new large-scale studies that link prenatal maternal & child health with family conditions and cognitive outcomes; and performing new kinds of analysis of existing data sources. Probabilistic programming plays an important role, addressing two key challenges: (i) there is far more routine data cleaning, exploration, imputation, and predictive modeling work than can be practically performed by the >20 person quantitative staff and (ii) new data sources such as GUSTO and INTERGROWTH present methodological challenges (around high dimensions, heterogeneous sources, etc) that currently lack solutions.

This poster presents preliminary data sketches obtained by short MML and BQL programs that BMGF viewed as core validation that BayesDB and probabilistic programming can automate routine baseline modeling and data exploration, and produces results that are in accord with common-sense knowledge. It also shows preliminary evidence for predictive signals --- between micronutrients in the mother’s blood at week 26 of pregnancy; anthropometry at birth; and cognitive outcomes at month 24, as assessed by Bayley scores --- that may warrant further in-depth exploration.

GUSTO - Growing Up in Singapore Towards healthy Outcomes
Growing Up in Singapore Towards healthy Outcomes (GUSTO) is the largest birth cohort study in Singapore yet, aimed at studying how a mother’s diet and lifestyle during pregnancy impact the growth of her children. We trained 32 models for 100 iterations on a subset of the GUSTO datatable using Crosscat.

Hierarchical clustering of the dependence probability grouped together variables that:
• Had same measurement time (e.g., at Pregnancy Week 26)
• Belonged to Father or Mother
• Regarded anthropometric measurements of the toddlers
• Regarded psychological scores (Bayley or Beck) of the toddlers

Intergrowth - Interbio
INTERGROWTH-21st is a population-based multiethnic cohort aimed at developing international standards for fetal and postnatal growth based on a prescriptive approach. The study is taking place in eight nations, namely Brazil, China, India, Italy, Kenya, Oman, UK and USA, and it focus on collecting information regarding growth, health and nutrition from early pregnancy to infancy in healthy growth environments.

The INTERBIO-21st Study is an extension to the INTERGROWTH-21st project, aimed at improving characterization from preterm birth syndromes and intrauterine growth restriction. The data gathered encompasses standardised information on nutrition, pregnancy outcomes, newborn anthropometric measurements, and on growth and development up to 48 months of age.

Acknowledgements & References
The authors would like to thank Feras Saad, Gregory Martin and Taylor Campbell for helpful feedback and discussions. This research was supported by DARPA (PPAML program, contract number FA8750-14-2-0004), IARPA (under research contract 2015-15061000003), the Office of Naval Research (under research contract N000141310333), the Army Research Office (under agreement number W911NF-13-1-0212), and gifts from Analog Devices and Google.