ML and precision public health: Saving mothers and babies from dying in rural India

Kasey Morris[*]
Surgo Foundation
Washington DC, USA
kaseymorris@surgofoundation.org

Vincent S. Huang[*]
Surgo Foundation
Washington DC, USA
vincenthuang@surgofoundation.org

Mokshada Jain
Surgo Foundation
Washington DC, USA
mokshadajain@surgofoundation.org

B.M. Ramesh
University of Manitoba
Winnipeg MB, Canada
ramesh.banadakoppamanjappa@umanitoba.ca

Hannah Kemp
Surgo Foundation
Washington DC, USA
hannahkemp@surgofoundation.org

James Blanchard
University of Manitoba
Winnipeg MB, Canada
james_blanchard@umanitoba.ca

Shajy Isac[†]
University of Manitoba
Winnipeg MB, Canada
shajy.isac@ihat.in

Bidyut Sarkar[†]
University of Manitoba
Winnipeg MB, Canada
Bidyut.sarkar@ihat.in

Vikas Gothalwal[†]
University of Manitoba
Winnipeg MB, Canada
vikasgothalwal@gmail.com

Vasanthakumar Namasivayam
University of Manitoba
Winnipeg MB, Canada
drvasanth@ihat.in

Sema Sgaier[‡]
Surgo Foundation
Washington DC, USA
semasgaier@surgofoundation.org

Abstract

Ambitious global health goals coupled with limited resources call for more targeted interventions. The advent of new and better data collection techniques, as well as advanced analytic methods, enhance our ability to take a precision public health approach. Highlighting a use case for increasing hospital delivery among pregnant mothers in northern India, we demonstrate how optimized data collection, combined with integrated machine learning (ML) methodologies, can be leveraged to effectively design and target precision interventions. We used predictive modeling to identify a broad set of factors related to hospital delivery and causal ML to determine the cause and effect ordering of those factors. A supervised ML algorithm was used to model individual heterogeneity by segmenting mothers into distinct types, with differentiated characteristics, behaviors, and risk profiles. Taken together, these findings provide a holistic picture of the drivers and barriers

[*] Equal contribution; joint first author

[†] Additional affiliation: India Health Action Trust, Lucknow, Uttar Pradesh, India

[‡] Corresponding author; additional affiliations: Harvard T.H. Chan School of Public Health; University of Washington

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

to hospital delivery in rural India and demonstrate how ML can make precision public health a reality.

1 Introduction

Global health has ambitious goals for improving the health and wellbeing of people in low income settings, but limited resources call for a more targeted approach. In India, where maternal and neonatal mortality remains stubbornly high, broad, system-wide intervention programs were implemented with the aim of increasing hospital facility delivery among pregnant mothers. These programs were initially successful and hospital deliveries increased by 40% over a 10-year period (Joe et al., 2018). But in recent years, success has plateaued and roughly 20% of mothers continue to deliver their babies at home. To close that gap, we must take a precision public health (PxPH) approach — that is, targeting the right intervention, to the right person, at the right time and place (Khoury et al., 2016; Chowkwanyun et al, 2018). The greatest potential to invoke meaningful and sustainable change lies in the development of a holistic intervention portfolio, targeting multiple aspects of the PxPH framework, informed by multiple analytic methods. Integration of machine learning (ML) methods offers a promising path for making this approach a reality.

Here, we describe a use case for the application of ML to drive PxPH for increasing hospital facility delivery in Uttar Pradesh, India. We begin by focusing on the optimization of data collection. Standard data collection focuses on the *what* (i.e., behavior) instead of the *why* (i.e., drivers of behavior). Without the *why*, interventions are unlikely to result in impactful and sustainable behavior change. To that end, we designed a survey to systematically interrogate the drivers and barriers of hospital delivery (Engl et al., 2019). We then applied a multi-pronged analytic approach to generate insights: We used predictive modeling to find the factors that were associated with hospital delivery and applied causal ML to determine the relative importance and underlying causal structure of those factors. We then used a supervised ML algorithm to segment types of mothers and the unique set of factors that influence where they deliver their baby and why. We discuss insights generated from these analyses, lessons learned, and future applications of this work.

2 Methods

Data was collected from 5,968 mothers in Uttar Pradesh, India who had recently given birth. The community healthcare worker (Accredited Social Health Activists; ASHA) catchment area was the primary sampling unit. Data was collected from 75 Districts and 600 Blocks. The survey consisted of items assessing health behaviors during pregnancy, including antenatal care (ANC) checkups and use of iron and folic acid (IFA) supplements; frequency of visits from the ASHA; birth planning; opinion of available maternity health services and infrastructure; risk perception; finances (e.g., money borrowed for delivery); awareness of government financial incentives; and sociodemographics.

The primary outcome variable for all analyses was whether a mother delivered in a facility (public or private) or at home. From the 41 variables included in the regression model, 16 significantly increased the likelihood of delivering in a hospital. These factors were used in the causal ML analysis. The supervised ML segmentation was conducted on the full set of variables used in the regression model.

2.1 Analyses

2.2 Causal machine learning

We used Bayesian network for causal machine learning (Greenland et al., 1999; Pearl, 1995). Bayesian networks are probabilistic graphical models that represent the conditional dependencies underlying a set of variables. Bayesian networks leverage these conditional dependencies to model causation. The underlying causal ordering of factors is identified through the structural output (i.e., a directed acyclic graph [DAG], which represents variables and their 'edges,' or the directed paths between variables). This shows which variables are directly causal of the outcome of interest, which are causal through upstream pathways, and which are outside the causal chain.

We used GNS Healthcare's Reverse Engineering and Forward Simulation (REFS^{TM}) platform to generate our causal Bayesian networks. REFS uses a Markov Chain Monte Carlo (MCMC) method to build an ensemble of causal models. To improve computational efficiency, all continuous variables were converted to categorical and number of categories were reduced to 3 or less. Expert opinion was incorporated into the constraints of the structure learning process.

2.3 Segmentation using supervised machine learning

To model individual level heterogeneity, we used a chi-square automatic interaction detection analysis (CHAID) decision tree algorithm (Kass, 1980). We employed a top-down pruning approach by sequentially modifying the stopping criteria (minimum number of cases per node, maximum tree depth, and alpha threshold) to be more stringent. A 10-fold cross-validation method was used to evaluate the model. A total of five trees were examined; the tree with the simplest structure and lowest prediction and generalization error was chosen as the final model. After the final tree was constructed, the segment identifier for each case was saved and the segments were profiled on the set variables to determine characteristics, drivers of behavior, and risk profiles.

3 Results

3.1 Bayesian networks identify causal factors driving hospital delivery

Figure 1 shows the consensus causal model built by REFS. Several variables are directly causal of delivery location: the perception that hospital delivery is safer than home, having a pre-determined delivery plan, being aware of financial incentives, education level, and being a first-time parent. More upstream, the number of ANC checkups is an important ‘gateway’ variable that is central to many causal pathways in the network. For example, increasing ANC checkups leads to positive opinions about the safety of hospital delivery, increased knowledge and awareness of the health services (and incentives), and committed behaviors (e.g., delivery plan). Additionally, education of the mother and perceiving hospital delivery as the social norm appear to be important internal and external causes, respectively, that modulate other downstream behaviors and perceptual opinions.

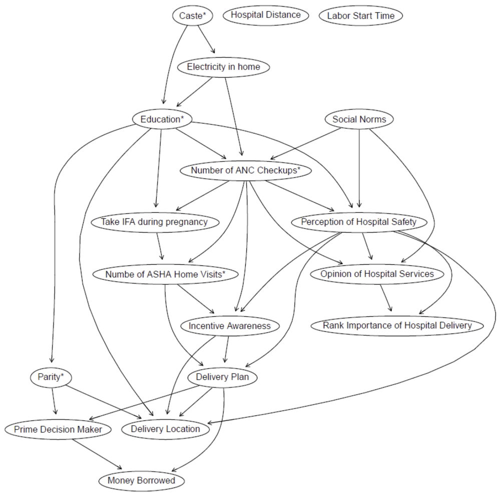

Figure 1: Structural graph depicting the causal relationships.

Although distance to the nearest hospital and the time of labor were both associated with delivery location they are not causal of it. Similarly, the primary decision maker of delivery location, whether the family borrowed money, and general opinions about hospital services were not causal of delivery location. Instead, they appear to be conflated with hospital delivery due to some common upstream causes (such as delivery planning).

3.2 Four segments can explain individual-level heterogeneity

The segmentation solution consisted of four types of mothers, based on three factors that split the mothers into subgroups (see Figure 2). The first split was made on perceptions of hospital safety, (χ^2 = 602.87, p < .001). Among mothers who believe the hospital is safer, the next split was made on delivery planning (χ^2 = 604.47, p < .001). Of the 26.3% of mothers who report their delivery location was a last-minute decision, 68.4% (n = 1049) delivered in a hospital. Of the 55.7% of mothers who delivered in their planned location, 94.7% (n = 3080) delivered in a hospital. Among mothers who believe the home is safer than the hospital, the next split was made based on parity (χ^2 = 84.07, p < .001). Of the 12.4% of experienced mothers in this group, 43.4% (n = 313) delivered in a hospital. Of the 5.6% of first-time mothers in this group, 73.9% (n = 243) delivered in a hospital.

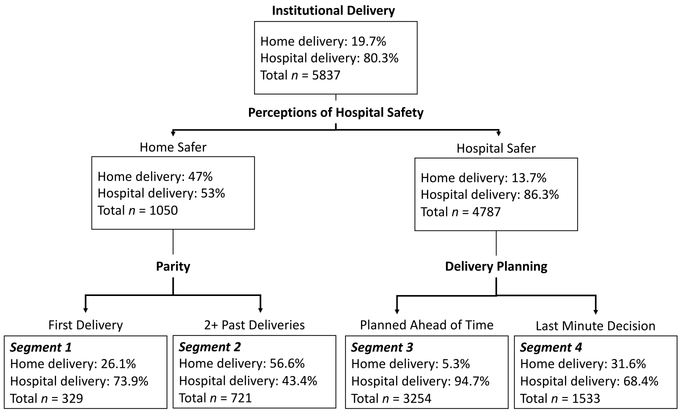

Figure 2: Final decision tree; values within nodes indicate the percent of each subgroup delivering at home or at a hospital facility and the total sample represented within the node. Note: The model correctly classified 81.9% of cases. The generalization risk estimate was .181 (SE = .005), indicating the model performed comparably on the validation samples.

Home delivery risk was well-differentiated by the segments, with 77% of all home deliveries occurring in just two segments (2 and 4). The home deliveries that occurred within Segment 4 were almost entirely due to incidental, or non-elective reasons (83.6%). For example, the most common reason given among these mothers who delivered at home was that the baby came too quick (54.9%). In contrast, more than half (56.2%) of all elective home deliveries are accounted for by Segment 2; for these mothers, the most common reason given for home delivery was that it was more convenient (36.3%).

4 Lessons learned, limitations, and future work

Two key lessons were learned in the application of a PxPH approach for increasing hospital facility delivery in rural India. First, this approach requires optimized data collection to encompass a broad set of drivers and barriers to behavior. ML methods generally follow a 'garbage in, garbage out' principle. Without this kind of optimized data, insights and associated intervention recommendations are likely to fall short. Second, the application of multiple ML algorithms is necessary to provide a 360-degree view of the outcome of interest. The ML methodologies used here range from well understood and refined algorithms (i.e., supervised learning) to novel and emerging approaches (i.e., causal ML). We show how each provides a new layer of information, and when considered together, allow for a precision, targeted approach.

This approach has several limitations to note. Although we can offer recommendations based on our insights, we cannot say, without evaluation metrics, whether those recommendations would result in meaningful change on the ground. Additionally, the insights derived from these ML methods are, by definition, dependent on the data that feeds into them. A lack of complete or high-quality data could result in precision targeting recommendations that are incomplete, inaccurate, or biased. Careful collaboration between researchers, policy makers, and implementers is required to mitigate these risks.

We have applied this approach to two other use cases in low income settings, demonstrating the generalizability of adopting a PxPH framework.

4.1 Tuberculosis care seeking in Chennai, India

One of the easiest ways to reduce the spread of tuberculosis (TB) is to initiate medical intervention very early after the onset of symptoms. In Chennai, India we have completed a longitudinal study examining care seeking behavior among TB symptomatic individuals. We first used logistic regression to identify predictors associated with care-seeking for TB symptoms. We then used an unsupervised ML algorithm to segment individuals on those predictors and created a typing tool that identifies individuals who are particularly at risk of not seeking care for their symptoms.

4.2 Sexual and reproductive health in Madhya Pradesh, India

In a large-scale study on sexual and reproductive health behaviors in Madhya Pradesh, India, we plan to use predictive modeling and segmentation to uncover differences in household barriers towards using methods of family planning. As a large number of predictors are likely to influence behaviors around planning family size, we also plan to supplement predictive modeling with causal models to identify the most promising targets for intervention design.

References

[1] Chowkwanyun, M., Bayer, R., & Galea, S. (2018). “Precision” public health—between novelty and hype. New England Journal of Medicine, 379(15), 1398-1400.

[2] Engl E Sgaier SK. (2019). CUBES: A practical toolkit to measure enablers and barriers to behavior for effective intervention design [version 1; peer review: 2 approved, 1 approved with reservations]. Gates Open Res 2019, 3:886 (https://doi.org/10.12688/gatesopenres.12923.1)

[3] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.

[4] Joe, W., Perkins, J. M., Kumar, S., Rajpal, S., Subramanian, S. V. (2018). Institutional delivery in India, 2004–14: unravelling the equity-enhancing contributions of the public sector. Health policy and planning, 33(5), 645-653.

[5] Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat. 1980;29(2):119-127.

[6] Khoury, M. J., Iademarco, M. F., & Riley, W. T. (2016). Precision public health for the era of precision medicine. American journal of preventive medicine, 50(3), 398.

[7] Pearl J. Causal Diagrams for Empirical Research. Biometrika. 1995;82(4):669-88.

5 Supplementary Material

5.1 Causal ML: Interventional Query

Causal ML can be used to conduct intervention query or "what-if" analyses. This quantifies the change in the outcome variable that occurs when changing the level of a given input variable. We did this for all variables, essentially conducting a series of what-if analyses for the delivery location outcome. The results are plotted as odds ratios for hospital delivery.

We found that, by far, having delivery plan is the most influential cause of hospital delivery; mothers are more than six times likely to deliver in public facilities if there was a delivery plan than if there was not. Almost just as causal is the safety perception of hospital delivery (greater than 5 times). This result suggests key interventional areas that program may focus their resources on. Incentive awareness and mother's education are also important (greater than 3 times). Next is whether the woman is a first-time parent (2 times). The number of ANC checkups, home visits, and perceptions of hospital delivery being a social norm are also significant causes with more moderate levels.

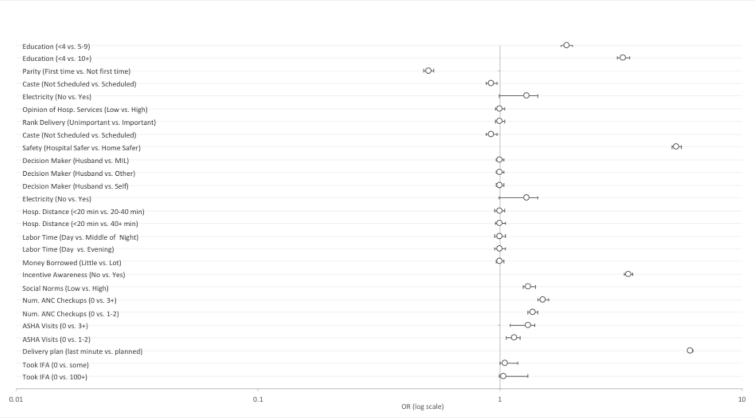

Figure 3: What-if analysis (reference/intervention); interventional odds ratio of hospital facility vs. home delivery with 95% confidence intervals.