Causal Inference - project proposal
Breastfeeding and Infant Mortality

Noam Sasson - 214439929 & Alon Eitan - 214304693

August 25, 2025

1 Causal Question

The causal question is: *Does breastfeeding during early stages of life affect infant mortality during the first year of life?*

2 Previous Knowledge On the Topic

1. Time to initiation of breastfeeding and neonatal mortality and morbidity: a systematic review (Smith et al. [2013]) - reporting a direct association between early breastfeeding initiation and neonatal mortality and morbidity outcomes.

2. Delayed breastfeeding initiation and infant survival: A systematic review and meta-analysis (Smith et al. [2017]) - Compared to infants who initiated breastfeeding $\leq$ 1 hour after birth, infants who initiated breastfeeding 2-23 hours after birth had a 33% greater risk of neonatal mortality.

3. Infant and young child feeding, a WHO fact sheet (World Health Organization [2021]) - Highly recommends early initiation of breastfeeding within the first hour of life.

3 Data We Intend To Use

We chose the 2015 Cohort Linked Birth/Infant Death Data Set from the site of the **National Center for Health Statistics**. The data contains records of all babies with birth certificates born in the United States In the year 2015, and their survival status for their first year (whether they died within 2015 or 2016). All the enteries of each record are the birth certificates that were filled on the birth of the baby, linked with the infant death certificates that were filled on the death of the baby (only if baby died), with additional questions that the patients were asked.

"The National Center for Health Statistics is the nation's source for official health statistics. We collect, analyze, and share data and statistics to guide programs and policies that improve the health of people across the United States."

4 Causal Assumptions

The infant mortality dataset contains a lot of information about the mother's and father's background and status, and the infant's health.
From the vast amount of features to select from, we collected $\sim$ 100 features spanning over 12 categories, which we believe are relevant to the causal question.
1. Maternal Demographics, **2.** Father Paternal Characteristics, **3.** Maternal Health & Risk Factors, **4.** Prenatal Care, **5.** Smoking / Tobacco Use, **6.** Anthropometrics / Weight, **7.** Pregnancy History, **8.** Infections, **9.** Labor & Delivery, **10.** Infant & Birth Outcomes, **11.** Congenital Anomalies of the Newborn, **12.** Maternal Mortality Features

Table 1: Features Dictionary for Infant Mortality Analysis

Maternal Demographics (mother)	**Paternal Characteristics (father)**
MAGER - Mother's Age Recode 41	FAGECOMB - Father's Combined Age (Revised)
MRACE31 - Mother's Race Recode 31	FRACE31 - Father's Race Recode 31
MEDUC - Mother's Education	FEDUC - Father's Education
DMAR - Marital Status	
MBSTATE_REC - Mother's Nativity	
RESTATUS - Residence Status	
WIC - participation in WIC program	
PAY - Payment Source	

Maternal Health & Risk Factors	**Prenatal Care**
RF_PDIAB - Pre-pregnancy Diabetes	PRECARE - Month Prenatal Care
RF_GDIAB - Gestational Diabetes	PREVIS - Number of Prenatal Visits (Revised)
RF_PHYPE - Pre-pregnancy Hypertension	
RF_GHYPE - Gestational Hypertension	
RF_EHYPE - Hypertension Eclampsia	
RF_PPB - Previous Preterm Birth	
RF_INFT - Infertility Treatment	
RF_DRG - Fertility Enhancing Drugs	
RF_ART - Assisted Reproductive Technology	
RF_CESAR - Previous Cesareans	
RF_CESARN - Number of Previous Cesareans	
NO_RISKS - No Risk Factors Checked	

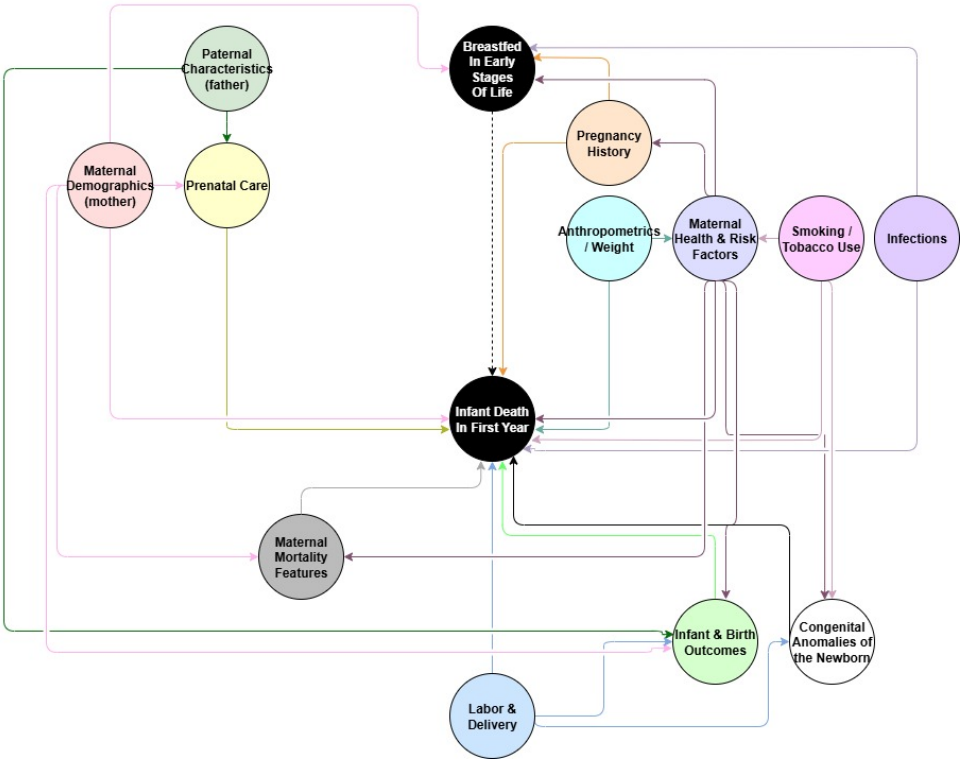
Smoking / Tobacco Use	**Anthropometrics / Weight**
CIG_0 - Cigarettes Before Pregnancy Recode	MHTR - Mother's Height in Inches (Recode)
CIG_1 - Cigarettes 1st Trimester Recode	BMI - BMI
CIG_2 - Cigarettes 2nd Trimester Recode	BMI_R - Body Mass Index Recode
CIG_3 - Cigarettes 3rd Trimester Recode	PWgt_R - Pre-pregnancy Weight Recode
CIG_REC - Cigarette Recode (Revised)	DWgt_R - Delivery Weight Recode
	WTGAIN - Weight Gain
	WTGAIN_REC - Weight Gain Recode

Pregnancy History	Infections
LBO_REC - Live Birth Order Recode	IP_GON - Gonorrhea
TBO_REC - Total Birth Order Recode	IP_SYPH - Syphilis
PRIORLIVE - Prior Births Now Living	IP_CHLAM - Chlamydia
PRIORDEAD - Prior Births Now Dead	IP_HEPB - Hepatitis B
PRIORTERM - Prior Other Terminations	IP_HEPC - Hepatitis C
ILLB_R - Interval of Last Live Birth Recode	NO_INFEC - No Infections Checked
ILLB_R11 - Interval of Last Live Birth Recode 11	
ILOP_R - Interval of Last Other Pregnancy Recode	
ILOP_R11 - Interval of Last Other Pregnancy Recode 11	
ILP_R - Interval of Last Pregnancy Recode	
ILP_R11 - Interval of Last Pregnancy Recode 11	

Labor & Delivery	Infant & Birth Outcomes
LD_INDL - Induction of Labor	SEX - sex of infant
LD_AUGM - Augmentation of Labor	COMBGEST - Combined Gestation Imputed
LD_CHOR - Chorioamnionitis	GESTREC10 - Combined Gestation Recode 10
LD_STER - Steroids	GESTREC3 - Combined Gestation Recode 3
LD_ANTB - Antibiotics	BWTR14 - Birth Weight Recode 14
LD_ANES - Anesthesia	APGAR5 - Five Minute APGAR Score
ME_PRES - Fetal Presentation	APGAR10 - Ten Minute APGAR Score
ME_ROUT - Final Route & Method of Delivery	DPLURAL - Plurality Recode
ME_TRIAL - Trial of Labor Attempted	SETORDER_R - Set Order Recode
DMETH_REC - Delivery Method Recode	AB_AVEN1 - Assisted Ventilation
RDMETH_REC - Delivery Method Recode Combined	AB_AVEN6 - Assisted Ventilation >6 hrs
	AB_NICU - Newborn Admitted to NICU
	AB_SURF - Newborn Received Surfactant Therapy
	AB_ANTI - Newborn Received Antibiotics
	AB_SEIZ - Newborn Experienced Seizures
	NO_ABNORM - No Abnormal Conditions Checked

Congenital Anomalies of the Newborn	Maternal Mortality Features
CA_ANEN - Anencephaly	MM_MTR - Maternal Transfusion
CA_MNSB - Meningomyelocele / Spina Bifida	MM_PLAC - Perineal Laceration
CA_CCHD - Cyanotic Congenital Heart Disease	MM_RUPT - Ruptured Uterus
CA_CDH - Congenital Diaphragmatic Hernia	MM_UHYST - Unplanned Hysterectomy
CA_OMPH - Omphalocele	MM_AICU - Admission to Intensive Care
CA_GAST - Gastroschisis	NO_MMORB - No Maternal Morbidity Checked
CA_LIMB - Limb Reduction Defect	
CA_CLEFT - Cleft Lip with or without Cleft Palate	
CA_CLPAL - Cleft Palate Alone	
CA_DOWN - Down Syndrome	
CA_DISOR - Suspected Chromosomal Disorder	
CA_HYPO - Hypospadias	
NO_CONGEN - No Congenital Anomalies Checked	

The causal assumptions we intend to use are mapped in the following causal DAG:



5 Challenges

1. Confounding Variables

   The primary challenge is to isolate the effect of early breastfeeding from numerous confounding variables that influence both the mother's decision to breastfeed and the infant's health outcomes. The project's extensive feature list and causal DAG highlight many of these factors, including:

   - Socioeconomic and Demographic Factors: Maternal demographics like age (MAGER), race (MRACE), and education (MEDUC) are likely correlated with both breastfeeding and infant mortality. Paternal characteristics, such as age (FAGECOMB) and education (FEDUC), are also included as potential confounders.
   - Maternal Health and History: The mother's pre-existing health conditions, such as diabetes (RF_PDIAB) or hypertension (RF_PHYPE), her pregnancy history (e.g., previous preterm births), and health during pregnancy (e.g., gestational diabetes) are significant confounders. These factors can affect a mother's ability to breastfeed and are also directly related to infant health.
   - Prenatal Care: The number of prenatal visits (PREVIS) and the month prenatal care began (PRECARE) are key confounders as they reflect a mother's access to and engagement with healthcare, which can influence both early breastfeeding support and infant outcomes.
   - Smoking/Tobacco Use: The amount of cigarettes smoked before and during pregnancy (CIG_0, CIG_1, CIG_2, CIG_3) is a known risk factor for poor infant

health and may be associated with breastfeeding decisions, making it a critical confounder to control for.

2. Reverse Causality and Endogeneity

   The analysis may be affected by reverse causality, where the infant's health status influences the mother's ability or decision to breastfeed, rather than the other way around. For example, a newborn with a congenital anomaly (e.g., Down Syndrome or Cleft Palate) may have difficulty breastfeeding, and this condition also increases the risk of mortality. This relationship is depicted in the causal DAG, with an arrow from :Congenital Anomalies of the Newborn: to :Infant Death in first year:. Similarly, :Infant & Birth Outcomes: like low birth weight (BWTR14) and low Apgar scores (APGAR5) could affect breastfeeding. This endogeneity makes it challenging to disentangle the true causal effect of breastfeeding.

3. Causal Inference from Observational Data

   The project relies on an infant mortality dataset, which is a form of observational data. Because a randomized controlled trial is not possible for ethical reasons, establishing a causal link requires careful statistical methods to control for observed and unobserved confounding. While the provided project outlines a detailed plan to use a large number of features and a causal DAG to map relationships, there remains the risk of unobserved confounding variables that are not captured in the dataset. This could lead to biased estimates of the effect of breastfeeding.

6 Estimation Methods

Our goal is to estimate the ATE of breastfeeding on infant mortality during the first year of life (target variable is probability of death in the first year). We are going to use the following estimation methods shown in the course:

1. Matching: We will use matching techniques seen in the course (k-nearest neighbor matching and propensity score matching).

2. T-learner & S-Learner: specificaly using

   - Logistic Regression both trained with IPW to handle treatment label imbalance (IPW Derived from a learned propensity model).
   - Neural Networks (NN) with similar IPW training, or smart balanced sampling from both treatment and control groups during training.

7 Robustness Checks

To evaluate the reliability of our causal estimates, we will conduct a series of robustness checks, testing our matching, T-learner, and S-learner models.

For the matching approach, we will examine the stability of our findings by varying key parameters. This includes exploring different caliper widths, adjusting the number

of matches used for each treated unit, and employing alternative distance metrics, such as the Mahalanobis distance. Consistency in the estimated treatment effect across these variations will serve as a strong indicator of the model's stability.

For the meta-learners, including the T-learner and S-learner, a crucial robustness check will involve assessing the dependence of our results on the choice of the underlying base learners. We will compare the causal estimates obtained from different regularization functions for the Logistic Regression, and same for another learner such as Random Forest - consistent Results on the 3 models (Logistic Regression, Random Forest and NN) will indicate a significant ATE.

Together, these multifaceted checks will provide confidence that our causal estimates are robust to different modeling choices and potential sources of uncertainty.

References

Amanda K. Smith et al. Time to initiation of breastfeeding and neonatal mortality and morbidity: a systematic review. *BMC Public Health*, 13:19, 2013. doi: 10.1186/1471-2458-13-19. URL `https://pubmed.ncbi.nlm.nih.gov/24564770/`.

Emily R. Smith, Lisa Hurt, Ranadip Chowdhury, Bireshwar Sinha, Wafaie Fawzi, and Karen M. Edmond. Delayed breastfeeding initiation and infant survival: A systematic review and meta-analysis. *PLoS One*, 12(7):e0180722, 2017. doi: 10.1371/journal.pone.0180722. URL `https://pubmed.ncbi.nlm.nih.gov/28746353/`.

World Health Organization. Infant and young child feeding. WHO Fact Sheet, 2021. URL `https://www.who.int/news-room/fact-sheets/detail/infant-and-young-child-feeding`. Accessed: 2024.