Causal effect of Exclusive Breastfeeding over Chronic illness

Shahar Rotem, Ofir Danan, Afik Bar

Technion – Israel institute of Technology

Causal effect of Exclusive Breastfeeding over Chronic illness

Breastfeeding provides unsurpassed natural nutrition to the newborn and infant. Human breast milk also contains numerous protective factors against infectious disease and may influence immune system development. If immune system development is significantly improved with the introduction of components of breast milk, then prematurely discontinued breastfeeding may facilitate pathogenesis of many chronic diseases later in life.

This article examines the hypothesis that breastfeeding may have long-term protective effects against chronic disease in children. Alternatively, artificial feeding, or the absence of breastfeeding, may increase the risk for chronic disease.

The literature reporting investigations of the association between infant feeding practices and chronic disease has expanded dramatically since 1984 when Borch-Johnsen and colleagues reported that the risk of type I diabetes, or insulin-dependent diabetes mellitus (IDDM), was higher among children who were not breastfed for at least 3 months. In addition, other studies have considered whether artificial feeding increases the risk for celiac disease, inflammatory bowel disease, childhood cancer, obesity and more.

In our research, we will try to estimate the casual effect of exclusive breastfeeding over chronic diseases (any) in children, using data from national health survey in India.

Data

The data we will use to test our hypothesis is a demographic survey in India (Annual Health Survey 2012-2013), which describes key indicators extracted from nine states in India [*Assam, Bihar, Jharkhand, Uttar Pradesh, Uttarakhand, Madhya Pradesh, Chhattisgarh, Odisha, Rajasthan*].

The survey objective was to yield a comprehensive, representative and reliable data on vital indicators including composite ones like Infant Mortality Rate, Maternal Mortality Ratio and Total Fertility Rate along with their co-variates. However, we can use multiple features from the survey for our specific hypothesis.

The survey was conducted by the Ministry of Health and Family Welfare in India and covered 4.3 million households from 284 districts. The data is summarized on a district level.

We've collected this data from India's Open Government Data (OGD), which is an initiative of Government of India and the US government to increase transparency in and allow collaboration.

The data contains over 600 features, categorical and numerical, from 26 different key indicator motifs.

Since our casual effect hypothesis is a medical one, naturally – it is very hard to isolate its confounders (observed and hidden). Thus, our first challenge was selection and elimination of related features from the data. Our approach to this challenge was a minimal selection from different key indicators, i.e. cover multiple related (to our hypothesis) key indicators and from each a minimal set of features.

Following our casual question, we will assign treatment *[T]* as the percentage of exclusively breastfed infants for at least 6 months. We've decided to assign **3** different treatments: Low breastfeeding (0-20%), Medium (20-40%), High (40-100%). *[0, 1, 2 respectively]*

Our experiment outcome *[Y]* is the percentage of people exhibiting chronic illness symptoms. Our hypothesis is that there is a negative correlation between our treatment and outcome, i.e. as exclusive breastfeeding is more common, we expect lower rates of chronic illness symptoms.

Since the data is collected within small timeframe, there exists a time gap between the breastfed infants, and the population exhibiting chronic illness symptoms. However, we assume that behavioral patterns within each district are kept on average, thus we can expect same breastfeeding percentage over time. This assumption has substantial precedents, for instance Israelis exhibit higher tolerance to peanuts allergies than many countries. This is believed to be due to the **behavioral pattern** where mothers tend to feed peanut butter snack to infants.

All our confounders *[X]* in the experiment are **pre-treatment**.

- Population Density – Rural or Urban. This feature might greatly affect treatment and outcome. For instance, Urban areas tend to have higher air pollution, which might affect chronic illness percentages. Furthermore, Urban female population tend to work more commonly, which might lead to decreased breastfeeding habits.
- Literacy rate – We've used this feature as progression (social, academic, etc.) indicator for a district.
- Complicated medical operations performed – We've used this feature as healthcare status indicator. We assume that the more complicated operations performed, general healthcare is better.
- Ante Natal Care – The percentage of pregnant women who received pre-labor care.

Schema

Feature Name	Description	Type	Values
State_Name	State name	Categorical	Nine possible states
State_District_Name	District name	Categorical	284 districts
Is_Rural	Population density (Rural or Urban)	Boolean	True \ False
Sex_Ratio	Ratio between male and females in district.	Numerical	0-100
Literacy_Rate	Effective Literacy rate; Percentage of literates from population above age 7.	Numerical	0-100
Number_Of_Injured_Persons_Who_Received_Major_Treatment	Number of major treatments received.	Numerical	0-100
Number_Of_Injured_Persons_Who_Received_Severe_Treatment	Number of severe treatments received.	Numerical	0-100
Number_Of_Persons_Having_Any_Kind_Of_Symptoms_Of_Chronic_Illness	Chronic Illness symptoms frequency.	Numerical	0-100
Mothers_Who_Received_Any_Antenatal_Check_Up	Antenatal Checkup (Any) frequency.	Numerical	0-100
Institutional_Delivery	Birth given at trained healthcare professionals.	Numerical	0-100
Mothers_Who_Did_Not_Receive_Any_Post_Natal_Check_Up	No Postnatal Checkup rates.	Numerical	0-100
Children_Aged_12_23_Months_Fully_Immunized	Children (1-2 Years) fully immunized rate.	Numerical	0-100
Children_Who_Received_At_Least_One_Vitamin_A_Dose_During_Last_Six_Months	Children (6-35 Months) who received Vitamin A.	Numerical	0-100
Children_With_Birth_Weight_Less_Than_2_5_Kg	Birth weight lower than 2.5Kg rate.	Numerical	0-100
Children_Exclusively_Breastfed_For_At_Least_Six_Months	Children (6-35 Months) who were exclusively breastfed for at least six months.	Numerical	0-100

Preprocessing

Since each confounder might be divided by Population density (Urban or Rural), we've added a Boolean feature `Is_Rural` which indicates whether a record is Rural or Urban. This resulted in duplication of records and merging features that were split by population density.

We've also scaled multiple confounders that were normalized by some population quantity, to align all fields to 0-100 normalization (percentages).

To handle missing values, we've used mean imputation method, feature-wise.

Our initial data exploratory analysis revealed the variables distributions and pair-wise correlations (See appendix). The following chart visualize our treatment assignment groups size:

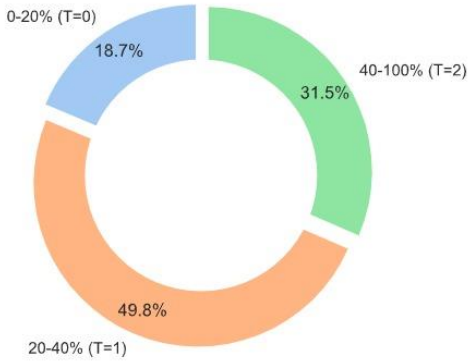


Figure 1. Treatment groups. Derived from exclusively breastfeeding percentage in district.

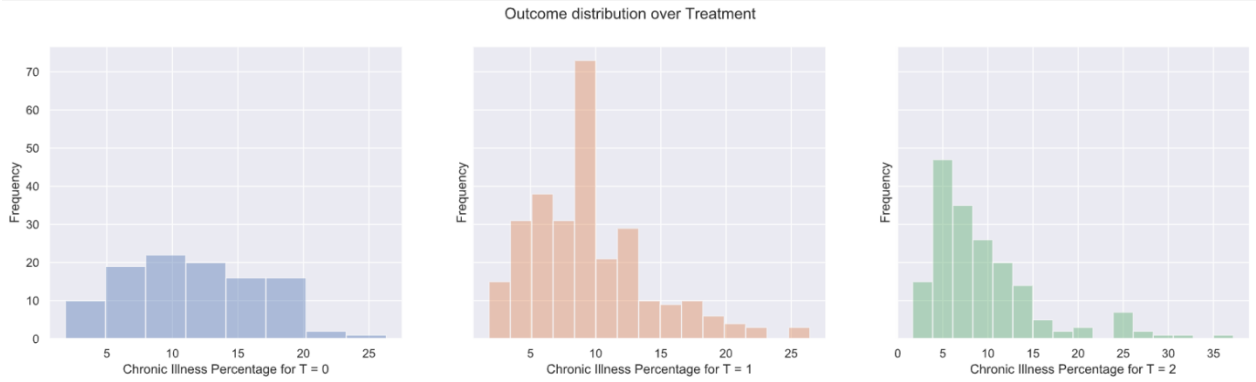


Figure 2. Chronic illness distribution by treatment group [Low, Medium, High]. We can see that as treatment increases (higher breastfeeding rate), the critical mass of chronic illness shifts to lower percentages.

The above correlation is not enough to determine causation; however, it provides some level of indication that there is **some** correlation. We tend to further investigate the existence of causation between the treatment and outcome.

Propensity Score Estimation

The propensity score of a subject is defined as function that receives the subject pretreatment covariates, x: $r(t,x) \coloneqq P(T = t \mid x), t \in \{1, \dots, M\}$. That is, the probability that subject with x as pretreatment covariates will receive treatment t.

In terms of comparison between two individuals, we need to compare a **vector** of propensity scores for each treatment: $R(x) = \big(r(1,x), r(2,x), \dots, r(M,x)\big)$.

We can estimate this probability function by training a Probabilistic classification model on our data. This model will be used to extract the probabilities to each treatment assignments.

We've experimented with two optional models: Multinomial Logistic Regression and Gradient Boosting Classifier. To compare them, we've used traditional model evaluation techniques – Train\Test split and ROC-curve analysis. The prevailing model was Gradient Boosting Classifier with 20 estimators.

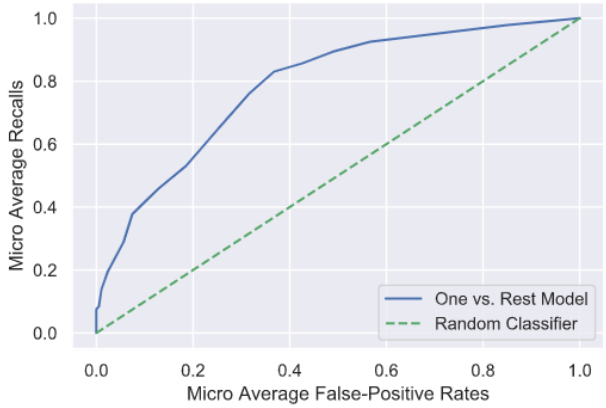


Figure 3. Gradient Boosting Classifier ROC-curve. In order to assess the quality of our predictor, we have generated a ROC curve on the test set. AUC = 0.78642, which indicates our model predicts treatments well.

We've used the selected model to extract propensity scores from our confounders.

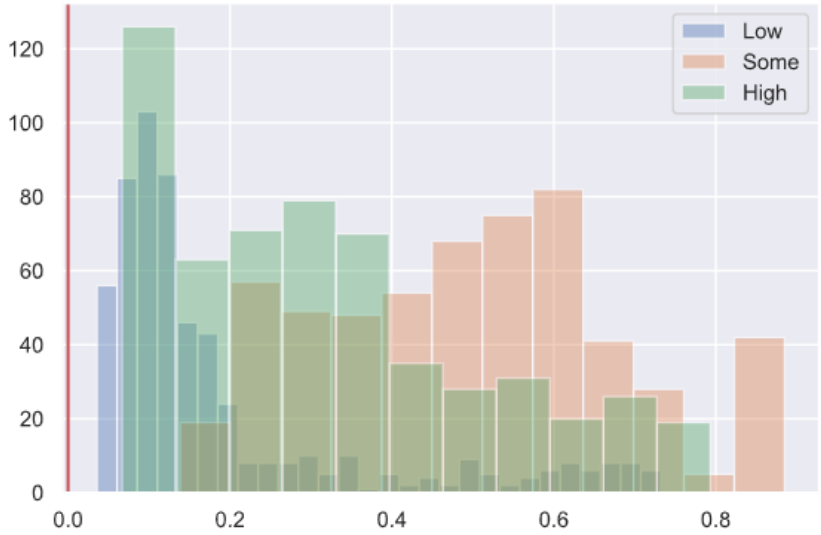


Figure 4. Propensity Scores from our model, by treatment groups. Low (0-20%), Some (20-40%), High (40-100%).

We've tried to improve the above treatment overlap and applied common support trimming. The left trimming bound was determine as the maximum of minimum values by treatment groups. Whereas the right trimming bound was the minimum of maximum values by treatment groups. Any record that had **at least** one propensity score (for any treatment assignment) outside of the common support bound, was filtered. Note that this might result in filtration within common support! Since a record might have one propensity score outside the bounds, but the other two treatment probabilities lay within.

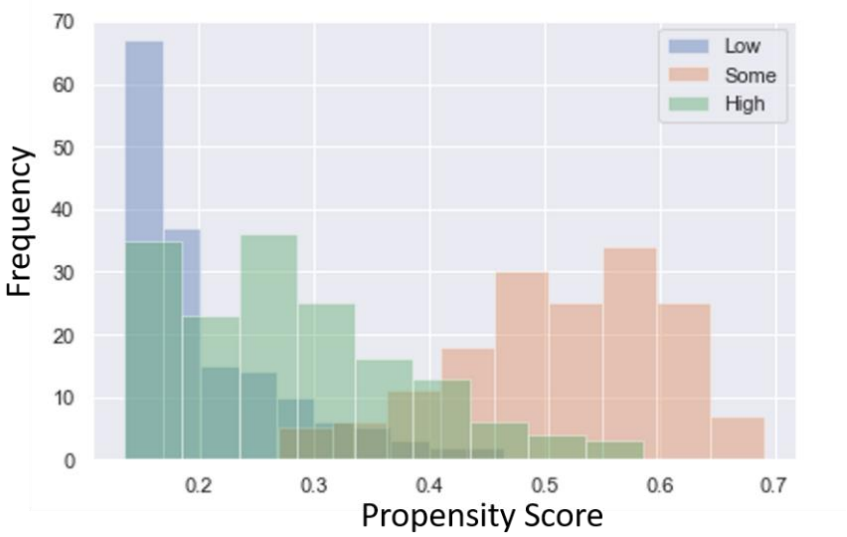


Figure 5. Propensity scores after common support trimming.

We can see that after common support trimming, we didn't obtain full overlap, however we did receive higher density and less data imbalance. Indeed, we've accomplished more coherent results with common support trimming and thus we've concluded that even though there is no full overlap, trimming did improve our data.

Causal Inference Justification

The data and our experiment don't represent a Randomized Control Trial. To assure causality we need to fulfill the following assumptions.

Common Support (Positivity) Assumption - $\forall t \in T, x \in X: P(T = t|X = x) > 0$. As we can see from our propensity score histogram, this assumption holds. However, we've performed the following verification:

```
# check overlap assumption (PS > 0):
np.all(propensity_score == 0)

False
```

Ignitability assumption - $\left((Y_0, Y_1, Y_2) \perp T\right) \mid X$. Thus, we need to verify there aren't any unmeasured confounders. It is extremely hard to fulfill this assumption as it is, since it's unlikely to find and measure all the confounders. However, we assume that the selected confounders from data will be enough for sufficient ignitability. We are aware of possible confounders which are not measured in this experiment, those will be listed under experiment limitations.

SUTVA (Stable Unit Treatment Value Assumption) –

Treatment applied to one unit doesn't affect the outcome for another unit; whether a person exhibits chronic illness symptoms doesn't affect another individual, regardless his treatment. However, there might be a genetic disturbance, which detail under limitations.

There is only a single version for each treatment level; we have 3 different treatments. Each treatment is clearly defined and distinguished from others. Even though each treatment group is consisted from a range of breastfeeding behaviors, we assume that small variations won't contributed greatly to the experiment outcome. Thus, each treatment is unique.

Consistency – Each district is assigned with a single treatment, and each district has its own chronic illness percentage. For consistency to hold, we need to assume that the **time gap** in our data doesn't affect treatment assignment. This will be explained in detail under limitations.

ATE Estimation Methods

To evaluate our treatment effect, we've used multiple methods to estimate ATE.

Inverse Propensity Weighting (IPW)

Inverse Propensity Weighting is a useful method to draw casual conclusions when treatment is not randomly assigned.

The propensity scores are unknown, and we estimated them with Gradient. as shown earlier.

We can use this quantity to normalize each group of treatment when computing average treatment effect, which will balance the study groups, and make them comparable.

To estimate ATE using IPW, we will use a pairwise ATE estimation (3 pairs).

$\forall(t_1, t_2) \in \{(0,1), (0,2), (1,2)\}$:

$$\widehat{ATE}_{(t_1,t_2)} = \left(\sum_{i=1}^{N} \frac{\mathbb{I}(T_i = t_1)Y_i}{r(t_1, X_i)}\right)\left(\sum_{i=1}^{N} \frac{\mathbb{I}(T_i = t_1)}{r(t_1, X_i)}\right)^{-1} - \left(\sum_{i=1}^{N} \frac{\mathbb{I}(T_i = t_2)Y_i}{r(t_2, X_i)}\right)\left(\sum_{i=1}^{N} \frac{\mathbb{I}(T_i = t_2)}{r(t_2, X_i)}\right)^{-1}$$

S-Learner

We've attempt to learn the correlations between X, T and Y. Thus, a Gradient Boosting Regressor model was used to estimate $Y \approx f(X, T)$.

This model was used to evaluate a pairwise ATE approximation:

$$\forall(t_1, t_2) \in \{(0,1), (0,2), (1,2)\}:$$

$$\widehat{ATE}_{(t_1,t_2)} = \frac{1}{N}\sum_{i=1}^{N} f(x_i, t_1) - f(x_i, t_2)$$

T-Learner

We've attempt to learn the correlations between X and Y subject to T. Thus, a Gradient Boosting Regressor models were used to estimate $Y_{t\in T} \approx f_{t\in T}(X)$, where each model was trained on his treatment group only. We've used these models to evaluate a pairwise ATE approximation:

$$\forall(t_1, t_2) \in \{(0,1),(0,2),(1,2)\}:$$

$$\widehat{ATE}_{(t_1,t_2)} = \frac{1}{N}\sum_{i=1}^{N} f_{t_1}(x_i) - f_{t_2}(x_i)$$

Results

Pair-wise treatment	IPW	S-Learner	T-Learner
(0, 1)	0.565623	0.225503	1.712349
(0, 2)	0.660937	0.225504	2.166147
(1, 2)	0.095314	0.0	0.453799

We can see that our ATE estimations are non-negative for all methods, which indicates a positive treatment effect. During our ATE estimations, we've compared a lower treatment to a higher treatment level (lower breastfeeding habits vs. higher breastfeeding habits). Thus, a positive ATE suggests that the outcome is higher for lower breastfeeding habits, which means **higher** chronically ill individuals in district. Thus, we can conclude a casual relation between breastfeeding for at least 6 months and chronic illness symptoms.

We can see that there isn't a drastic effect, at most 2% between the different treatment groups. However, if we will look at the distribution of our outcome, we will see that its critical mass is around 7% of entire population, so an effect of 2% is critical.

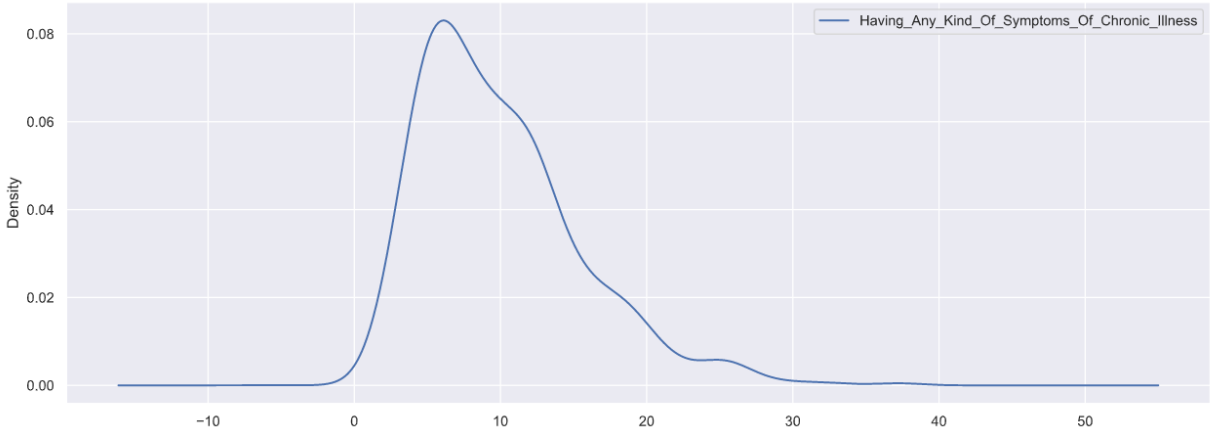


Figure 6. Outcome distribution.

In all our methods, the most dramatic effect was between $T = 0$ and $T = 2$, which aligns with our assumption that there is difference between lowest breastfeeding population and highest breastfeeding population, in terms of chronic illness frequency. Furthermore, we can see smaller effect between $T = 0$ and $T = 1$, $T = 1$ and $T = 2$, which aligns with our assumption that no small deviations in breastfeeding habits doesn't greatly influence outcome, and we can assume single treatment for each group.

Limitations and Discussion

The most crucial limitation in our experiment is hidden confounders. We can't measure their effect nor can map all possible hidden confounders that exists. Some hidden confounders are: Age, Lifestyle, Socio-economic status, genetic history and many more.

Summarized Data – The data is summarized by district level, which means we are using features of district to measure the casual effect. Furthermore, we do not have completely distinguishable control groups (data is not completely separable by treatments). The treatment is set by the majority of district. Same method is applied for Outcome.

Since the survey collected self-reported data during 2012-2013, there exist a time gap between the birth population and chronically ill population, thus we assume that the behavioral habits of population are kept, i.e. we can expect the same breastfeeding rate behavior in the past.

Our data is subjectable and may contain an inherent bias. All the data was collected only within India, which is a developing country, and suffer from multiple poverty influences. Furthermore, the nine states that the Annual Health Survey targeted, are high-focus states (which means, the poorest large states).

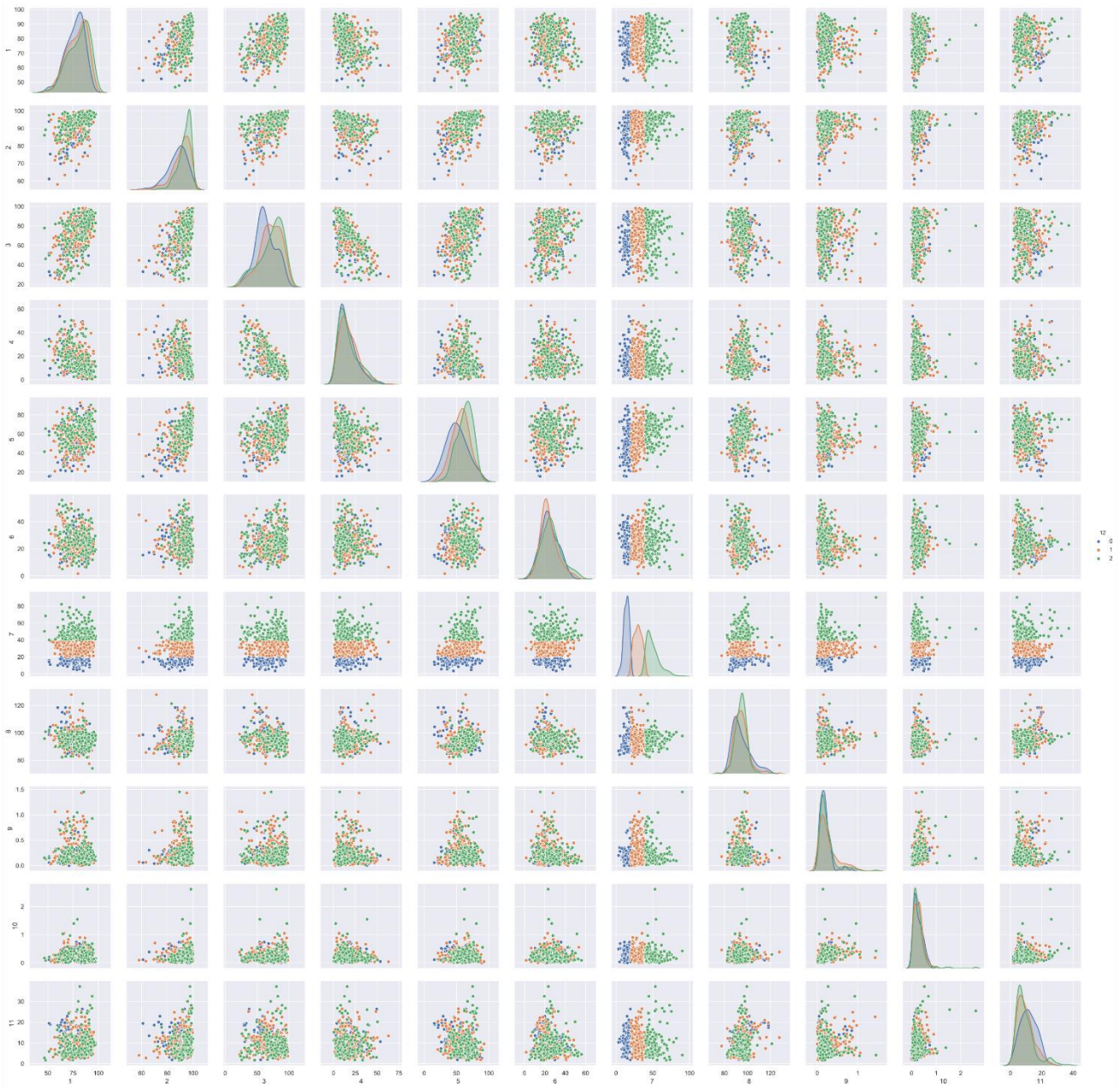
However, the alternative experiment is very unlikely – to track breastfeeding habits of individuals and follow their development is highly unlikely (time consuming and expensive).

In conclusion, we've learnt and experiment in multiple aspects during this project. We've resulted in a casual relation in our experiment which aligns with WHO (World Health Organization) breastfeeding recommendations.

Appendix

Variables pairwise distribution, hued by treatment group. Variable names are coded as:

1. Effective_Literacy_Rate
2. Mothers_Who_Received_Any_Antenatal_Check_Up
3. Institutional_Delivery
4. Mothers_Who_Did_Not_Receive_Any_Post_Natal_Check_Up
5. Children_Aged_6_35_Months_Who_Received_At_Least_One_Vitamin_A_Dose_During_Last_Six_Months
6. Children_With_Birth_Weight_Less_Than_2_5_Kg
7. Children_Aged_6_35_Months_Exclusively_Breastfed_For_At_Least_Six_Months
8. Sex_Ratio_All_Ages
9. Number_Of_Injured_Persons_By_Type_Of_Treatment_Received_Severe
10. Number_Of_Injured_Persons_By_Type_Of_Treatment_Received_Major
11. Having_Any_Kind_Of_Symptoms_Of_Chronic_Illness



References

Von Kries, R., Koletzko, B., Sauerwald et al, **Breast feeding and obesity: Cross sectional study**. Br Med J. 1999.

Borch-Johnsen, K., Joner, G., Mandrup-Poulsen, T. et al, **Relation between breast-feeding and incidence rates of insulin-dependent diabetes mellitus: A hypothesis**. Lancet. 1984.

American Academy of Pediatrics Work Group on Breastfeeding. **Breastfeeding and the use of human milk**. Pediatrics. 1997.

Journal of Allergy and Clinical Immunology. **Early consumption of peanuts in infancy is associated with a low prevalence of peanut allergy**. 2008.