Supplementary Data Description.
Below are the human-provided products for each of the 4 datasets A-D, as well as the research goals for datasets C and D.

A. Health Indicators dataset

General description of the dataset

The dataset includes diabetes related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), year 2015. The original BRFSS, from which this dataset is derived, is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

Data file description

"diabetes_binary_health_indicators_BRFSS2015.csv"
The csv file is a clean dataset of 253,680 responses (rows) and 22 features (columns).
All rows with missing values were removed from the original dataset; the current file contains no missing values.

The columns in the dataset are:

#1 `Diabetes_binary`: (int, bool) Diabetes (0=no, 1=yes)
#2 `HighBP`: (int, bool) High Blood Pressure (0=no, 1=yes)
#3 `HighChol`: (int, bool) High Cholesterol (0=no, 1=yes)
#4 `CholCheck`: (int, bool) Cholesterol check in 5 years (0=no, 1=yes)
#5 `BMI`: (int, numerical) Body Mass Index
#6 `Smoker`: (int, bool) (0=no, 1=yes)
#7 `Stroke`: (int, bool) Stroke (0=no, 1=yes)
#8 `HeartDiseaseorAttack': (int, bool) coronary heart disease (CHD) or myocardial infarction (MI), (0=no, 1=yes)
#9 `PhysActivity`: (int, bool) Physical Activity in past 30 days (0=no, 1=yes)
#10 `Fruits`: (int, bool) Consume one fruit or more each day (0=no, 1=yes)
#11 `Veggies`: (int, bool) Consume one Vegetable or more each day (0=no, 1=yes)
#12 `HvyAlcoholConsump` (int, bool) Heavy drinkers (0=no, 1=yes)
#13 `AnyHealthcare` (int, bool) Have any kind of health care coverage (0=no, 1=yes)
#14 `NoDocbcCost` (int, bool) Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? (0=no, 1=yes)
#15 `GenHlth` (int, ordinal) self-reported health (1=excellent, 2=very good, 3=good, 4=fair, 5=poor)
#16 `MentHlth` (int, ordinal) How many days during the past 30 days was your mental health not good? (1-30 days)
#17 `PhysHlth` (int, ordinal) Hor how many days during the past 30 days was your physical health not good? (1-30 days)
#18 `DiffWalk` (int, bool) Do you have serious difficulty walking or climbing stairs? (0=no, 1=yes)

#19 `Sex` (int, categorical) Sex (0=female, 1=male)
#20 `Age` (int, ordinal) Age, 13-level age category in intervals of 5 years (1=18-24, 2=25-29, ..., 12=75-79, 13=80 or older)
#21 `Education` (int, ordinal) Education level on a scale of 1-6 (1=Never attended school, 2=Elementary, 3=Some high school, 4=High school, 5=Some college, 6=College)
#22 `Income` (int, ordinal) Income scale on a scale of 1-8 (1=<=10K, 2=<=15K, 3=<=20K, 4=<=25K, 5=<=35K, 6=<=50K, 7=<=75K, 8=>75K)

B. Social Network dataset

General description of the dataset

* Rationale:
The dataset maps US Congress's Twitter interactions into a directed graph with social interactions (edges) among Congress members (nodes). Each member (node) is further characterized by three attributes: Represented State, Political Party, and Chamber, allowing analysis of the adjacency matrix structure, graph metrics and likelihood of interactions across these attributes.

* Data Collection and Network Construction:
Twitter data of members of the 117th US Congress, from both the House and the Senate, were harvested for a 4-month period, February 9 to June 9, 2022 (using the Twitter API). Members with fewer than 100 tweets were excluded from the network.

- `Nodes`. Nodes represent Congress members. Each node is designated an integer node ID (0, 1, 2, ...) which corresponds to a row in `congress_members.csv`, providing the member's Represented State, Political Party, and Chamber.

- `Edges`. A directed edge from node i to node j indicates that member i engaged with member j on Twitter at least once during the 4-month data-collection period. An engagement is defined as a tweet by member i that mentions member j's handle, or as retweets, quote tweets, or replies of i to a tweet by member j.

* Data analysis guidelines:
- Your analysis code should NOT create tables that include names of Congress members, or their Twitter handles.
- Your analysis code should NOT create tables that include names of States, or their two-letter abbreviations. The code may of course do statistical analysis of *properties* related to States, but should not single out specific states.

Data file description

"congress_members.csv"
A csv file of members of the 117th Congress, including their Twitter handles, Represented State, Party, and Chamber.
Data source: `https://pressgallery.house.gov/member-data/members-official-twitter-handles`.
Rows are ordered according to the node ID, starting at 0.

Fields:

`Handle`: Twitter handle (without `@`)
`State`: Categorical; Two-letter state abbreviation; including also: "DC", "PR", "VI", "AS", "GU", "MP".
`Party`: Categorical; Party affiliation ("D", "R", or "I")
`Chamber`: Categorical; The member's chamber ("House", "Senate")

“congress_edges.dat”
This file provides the interaction network between members of the 115th US Congress on Twitter.
Download and adapted from: `https://snap.stanford.edu/data/congress-twitter`

Each line contains two integers (i, j), indicating a directed edge from node ID i to node ID j, compatible with nx.read_edgelist('congress_edges.dat', create_using=nx.DiGraph()). An i->j edge indicates that Congress member i had at least one tweet engaging with Congress member j during the 4-month collection period.

C. Treatment Policy dataset

General description of the dataset

A change in Neonatal Resuscitation Program (NRP) guidelines occurred in 2015:
Pre-2015: Intubation and endotracheal suction was mandatory for all meconium-stained non-vigorous infants
Post-2015: Intubation and endotracheal suction was no longer mandatory; preference for less aggressive interventions based on response to initial resuscitation.

This single-center retrospective study compared Neonatal Intensive Care Unit (NICU) therapies and clinical outcomes of non-vigorous newborns for 117 deliveries pre-guideline implementation versus 106 deliveries post-guideline implementation.

Inclusion criteria included: birth through Meconium-Stained Amniotic Fluid (MSAF) of any consistency, gestational age of 35–42 weeks, and admission to the institution's NICU. Infants were excluded if there were major congenital malformations/anomalies present at birth.

File descriptions

"meconium_nicu_dataset_preprocessed_short.csv"
The dataset contains 44 columns:

`PrePost` (0=Pre, 1=Post) Delivery pre or post the new 2015 policy
`AGE` (int, in years) Maternal age
`GRAVIDA` (int) Gravidity
`PARA` (int) Parity
`HypertensiveDisorders` (1=Yes, 0=No) Gestational hypertensive disorder
`MaternalDiabetes` (1=Yes, 0=No) Gestational diabetes
`ModeDelivery` (Categorical) "VAGINAL" or "CS" (C. Section)
`FetalDistress` (1=Yes, 0=No)
`ProlongedRupture` (1=Yes, 0=No) Prolonged Rupture of Membranes
`Chorioamnionitis` (1=Yes, 0=No)
`Sepsis` (Categorical) Neonatal blood culture ("NO CULTURES", "NEG CULTURES", "POS CULTURES")
`GestationalAge` (float, numerical). in weeks.
`Gender` (Categorical) "M"/ "F"
`BirthWeight` (float, in KG)
`APGAR1` (int, 1-10) 1 minute APGAR score
`APGAR5` (int, 1-10) 5 minute APGAR score
`MeconiumConsistency` (categorical) "THICK" / "THIN"
`PPV` (1=Yes, 0=No) Positive Pressure Ventilation
`EndotrachealSuction` (1=Yes, 0=No) Whether endotracheal suctioning was performed
`MeconiumRecovered` (1=Yes, 0=No)
`CardiopulmonaryResuscitation` (1=Yes, 0=No)
`ReasonAdmission` (categorical) Neonate ICU admission reason. ("OTHER", "RESP" or "CHORIOAMNIONITIS")
`RespiratoryReasonAdmission` (1=Yes, 0=No)
`RespiratoryDistressSyndrome` (1=Yes, 0=No)

`TransientTachypnea` (1=Yes, 0=No)
`MeconiumAspirationSyndrome` (1=Yes, 0=No)
`OxygenTherapy` (1=Yes, 0=No)
`MechanicalVentilation` (1=Yes, 0=No)
`Surfactant` (1=Yes, 0=No) Surfactant inactivation
`Pneumothorax` (1=Yes, 0=No)
`AntibioticsDuration` (float, in days) Neonate treatment duration
`Breastfeeding` (1=Yes, 0=No) Breastfed at NICU
`LengthStay` (float, in days) Length of stay at NICU
`SNAPPE_II_SCORE` (int) 0–20 (mild), 21–40 (moderate), 41- (severe)

Human-provided Research goal

Research goal:
Examining the impact of guideline change on neonatal treatment and outcomes.

Hypothesis:
- Change in treatment policy lead to change in treatments.
- The change in treatment policy improved neonatal outcome, measured by duration of stay, apgar scores, etc.

D. Treatment Optimization dataset

General description of the dataset

Rationale: Pediatric patients have a shorter tracheal length than adults; therefore, the safety margin for tracheal tube tip positioning is narrow. Indeed, the tracheal tube tip is misplaced in 35%–50% of pediatric patients and can cause hypoxia, atelectasis, hypercarbia, pneumothorax, and even death.
Therefore, in pediatric patients who require mechanical ventilation, it is crucial to determine the Optimal Tracheal Tube Depth (defined here as `OTTD`, not an official term).

Note: For brevity, we introduce the term `OTTD` to refer to the "optimal tracheal tube depth". This is not an official term that can be found in the literature.

Existing methods: The gold standard to determine OTTD is by chest X-ray, which is time-consuming and requires radiation exposure.
Alternatively, formula-based models on patient features such as age and height are used to determine OTTD, but with limited success.

The provided dataset focus on patients aged 0-7 year old who received post-operative mechanical ventilation after undergoing surgery at Samsung Medical Center between January 2015 and December 2018.
For each of these patients, the dataset provides the OTTD determined by chest X-ray as well as features extracted from patient electronic health records.

File descriptions

"tracheal_tube_insertion.csv"
The csv file is a clean dataset of 969 rows (patients) and 6 columns:

Tube:
#1 `tube` - "tube ID", internal diameter of the tube (mm) [Included only for the formula-based model; Do not use as a machine-learning model feature]

Model features:
#2 `sex` - patient sex (0=female, 1=male)
#3 `age_c` - patient age (years, rounded to half years)
#4 `ht` - patient height (cm)
#5 `wt` - patient weight (kg)

Target:
#6 `tube_depth_G` - Optimal tracheal tube depth as determined by chest X-ray (in cm)

Human-provided Research goal

We formulated 6 different research goals that differ in the breadth of requested analysis and in the provision of mathematically explicit instruction.

Da: 4 Machine-Learning and 3 formula-based models

```
## Research Goal:

To construct and test 4 different machine-learning models and 3 different
formula-based models for the optimal tracheal tube depth (defined here as
`OTTD`, not an official term).

### ML MODELS:
Using the provided features (age, sex, height, weight), your analysis code
should create and evaluate the following 4 machine learning models for
predicting the OTTD:

- Random Forest (RF)
- Elastic Net (EN)
- Support Vector Machine (SVM)
- Neural Network (NN)

Important: It is necessary to hyper-parameter tune each of the models.

### FORMULA-BASED MODELS:
Your analysis code should compute the following 3 formula-based models for
the OTTD:

- Height Formula-based Model:
OTTD = height [cm] / 10 + 5 cm

- Age Formula-based Model:
optimal tube depth is provided for each age group:
0 <= age [years] < 0.5: OTTD = 9 cm
0.5 <= age [years] < 1: OTTD = 10 cm
1 < age [years] < 2: OTTD = 11 cm
2 < age [years]: OTTD = 12 cm + (age [years]) * 0.5 cm / year

- ID Formula-based Model:
OTTD (in cm) = 3 * (tube ID [mm]) * cm/mm

## Hypotheses:

- Each of the 4 machine learning models will have significantly better
predictive power than each of the formula-based models (as measured by their
squared residuals (prediction - target)**2 on the same test set).
```

Db: 1 Machine-Learning and 1 formula-based model

```
## Research Goal:
```

```
To construct and test 1 machine-learning model and 1 formula-based model for
the optimal tracheal tube depth (defined here as `OTTD`, not an official
term).

### ML MODEL:
Using the provided features (age, sex, height, weight), your analysis code
should create and evaluate the following 1 machine learning model for
predicting the OTTD:

- Random Forest (RF)

Important: It is necessary to hyper-parameter tune the model.

### FORMULA-BASED MODEL:
Your analysis code should compute the following 1 formula-based model for the
OTTD:

- Height Formula-based Model:
OTTD = height [cm] / 10 + 5 cm

## Hypothesis:

- The machine-learning model will have a significantly better predictive
power than the formula-based model (as measured by their squared residuals
(prediction - target)**2 on the same test set).
```

Dc: 2 Machine-Learning models

```
## Research Goal:

To construct and test 2 different machine-learning models for the optimal
tracheal tube depth (defined here as `OTTD`, not an official term).

### ML MODELS:
Using the provided features (age, sex, height, weight), your analysis code
should create and evaluate the following 2 machine learning models for
predicting the OTTD:

- Random Forest (RF)
- Elastic Net (EN)

Important: It is necessary to hyper-parameter tune each of the models.

## Hypothesis:
- The two machine-learning models will significantly differ in their
predictive power (as measured by their squared residuals (prediction -
target)**2 on the same test set).
```

Dai, Dbi and Dci:

We provided the same goals as above except with the omission of the explicit mathematical specification of the distance formula. Namely exactly the same goals, just deleting the specification: “`(prediction - target)**2`”.