

Extended Lombardy's Neonatal Screening Dataset

Received: 10 April 2025

Accepted: 27 November 2025

Cite this article as: Lopiano, G., Alberti, L., Cereda, C. *et al.* Extended Lombardy's Neonatal Screening Dataset. *Sci Data* (2025). <https://doi.org/10.1038/s41597-025-06389-4>

G. Lopiano, L. Alberti, C. Cereda & F. Cabitza

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

CONFIDENTIAL
COPY OF SUBMISSION FOR PEER REVIEW ONLY

Tracking no: SDATA-25-01828C

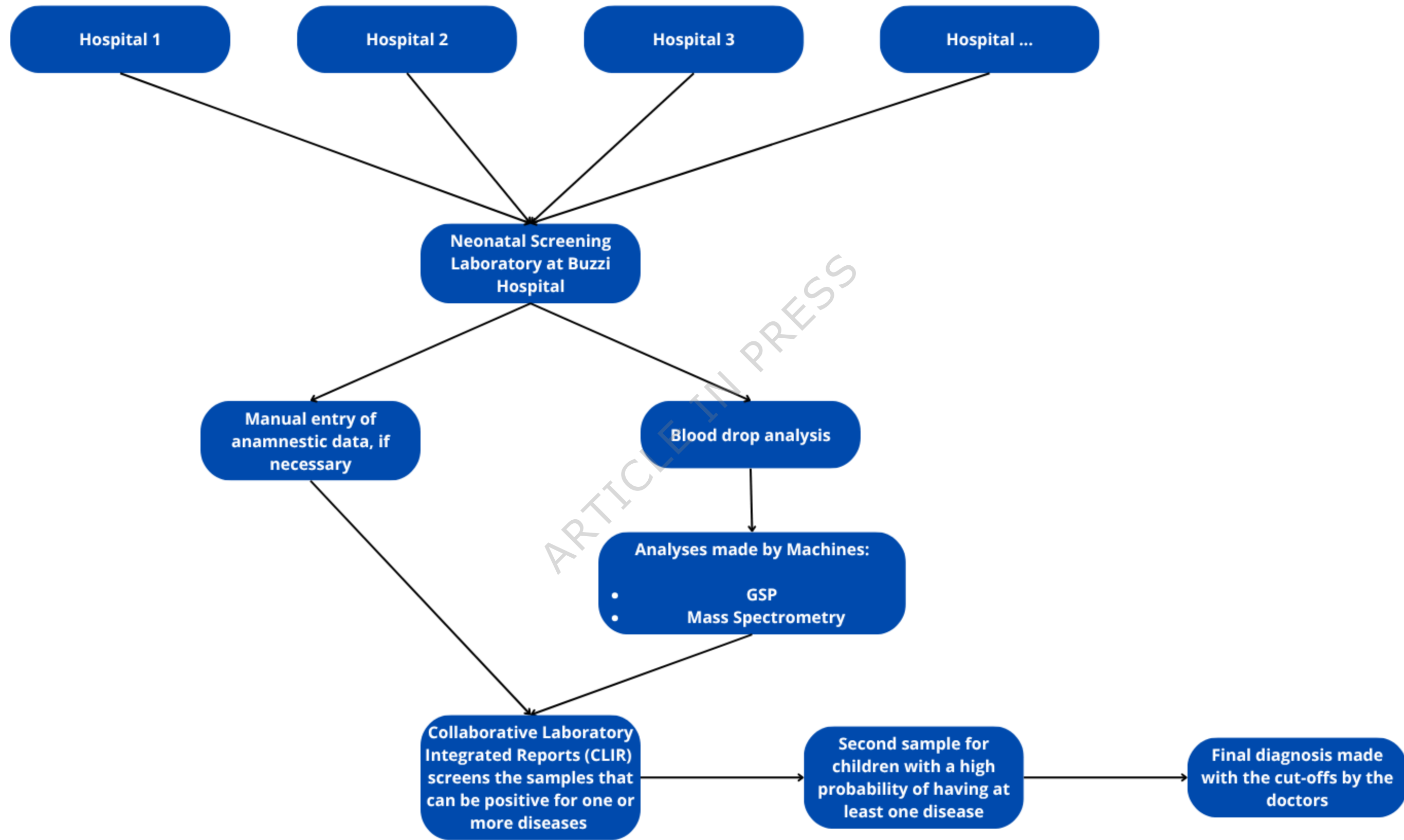
Extended Lombardy's Neonatal Screening Dataset

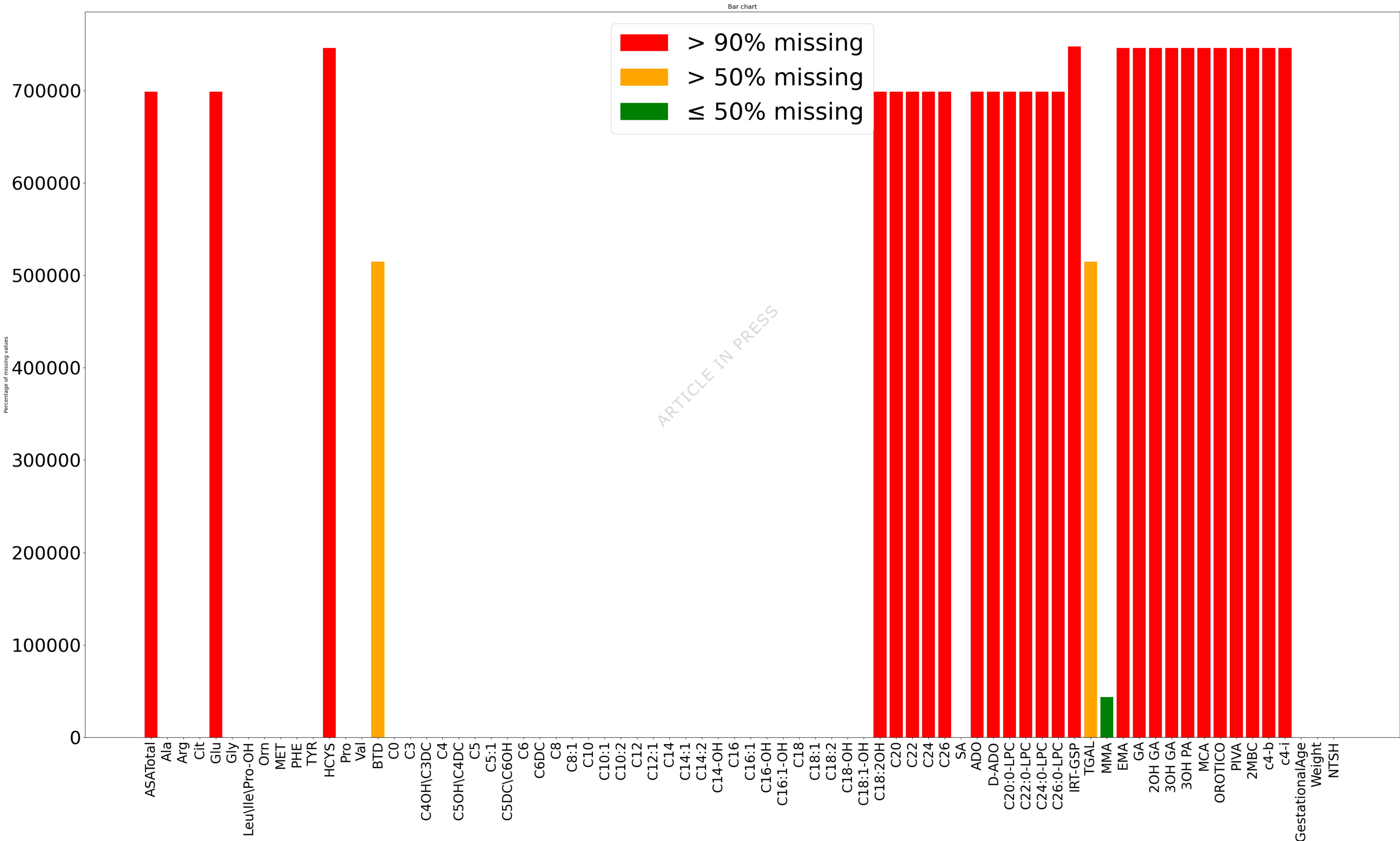
Authors: Gloria Lopiano (University of Milano Bicocca), Luisa Alberti (ASST Fatebenefratelli Sacco), Cristina Cereda (ASST Fatebenefratelli Sacco), and Federico Cabizza (Università di Milano-Bicocca)

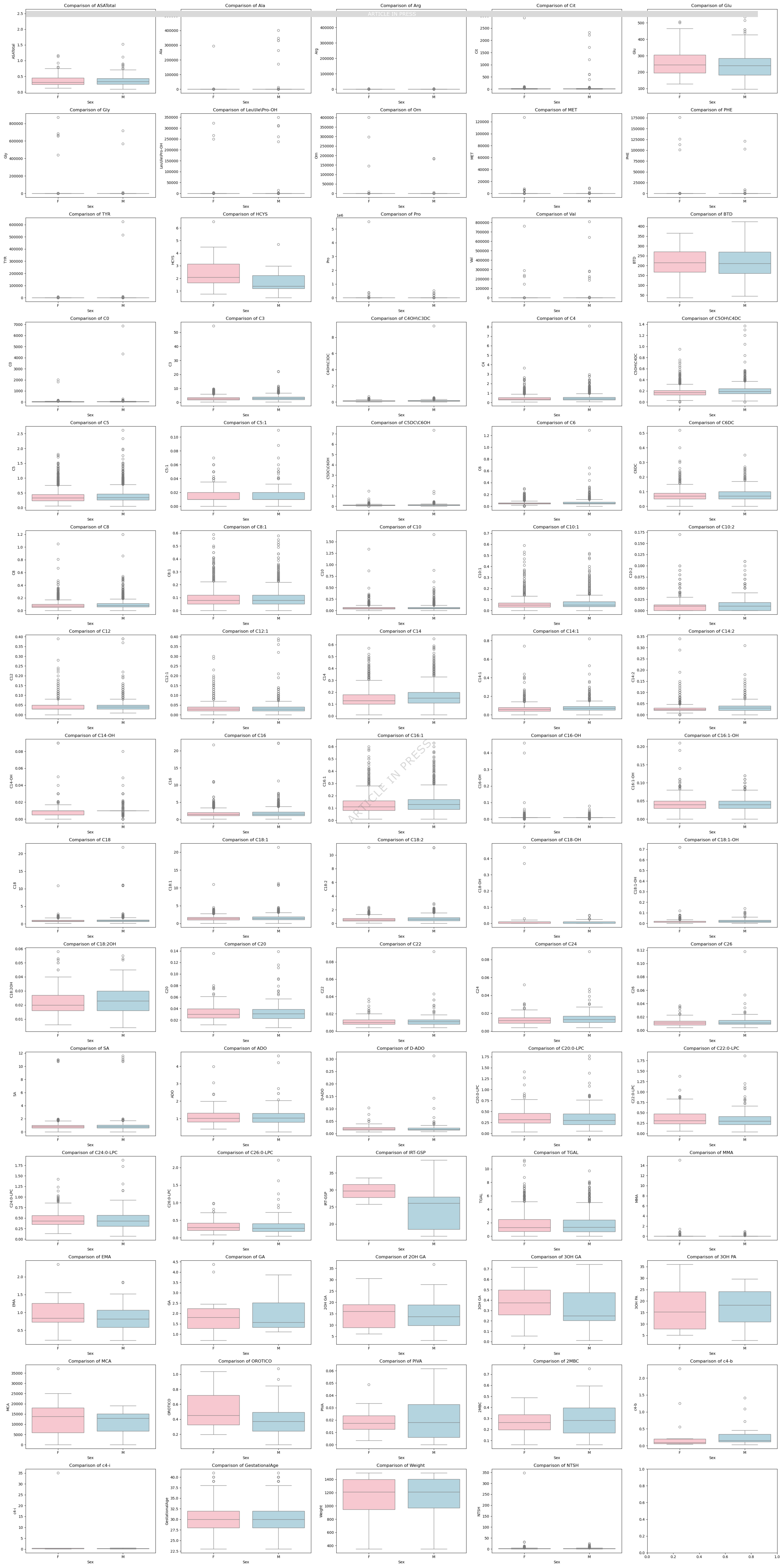
Abstract:
Neonatal Screening (NS) is a mandatory secondary prevention program established in the 1960s to detect rare genetic disorders in newborns. Early diagnosis through NS enables timely interventions, helping to prevent severe health complications. Despite its long-standing implementation, there remains a scarcity of publicly available datasets to support medical research and academic analysis. In this work, we present a comprehensive dataset containing blood samples and associated information from nearly a decade of Neonatal Screening in Lombardy. This dataset offers valuable insights for medical studies on genetic disorders, early diagnostics, and treatment efficacy. Additionally, it can support academic research in data-driven healthcare analysis and policy development. By making this dataset publicly available, we aim to enhance research opportunities in neonatal health and contribute to the improvement of early intervention strategies.

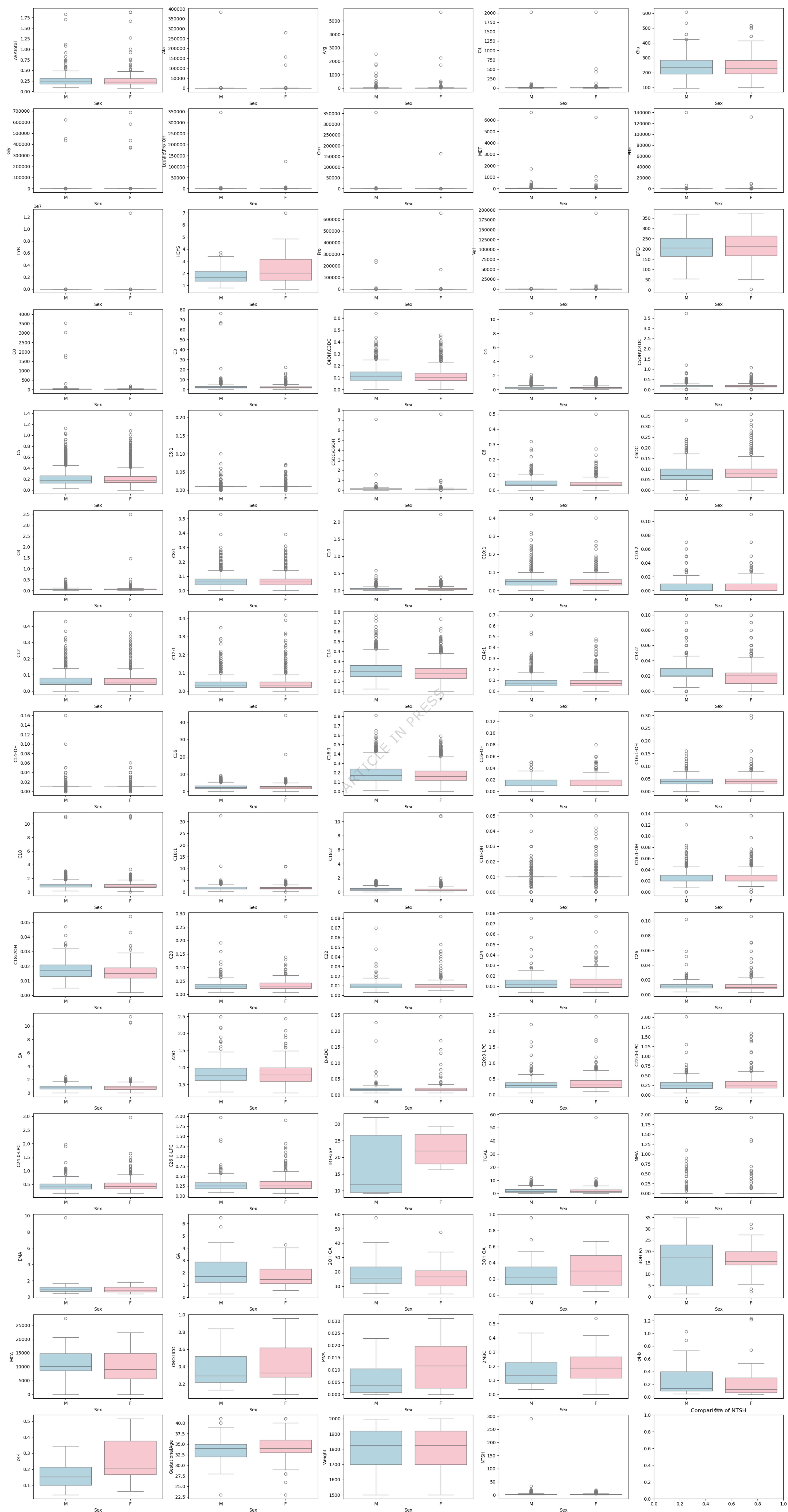
Datasets:

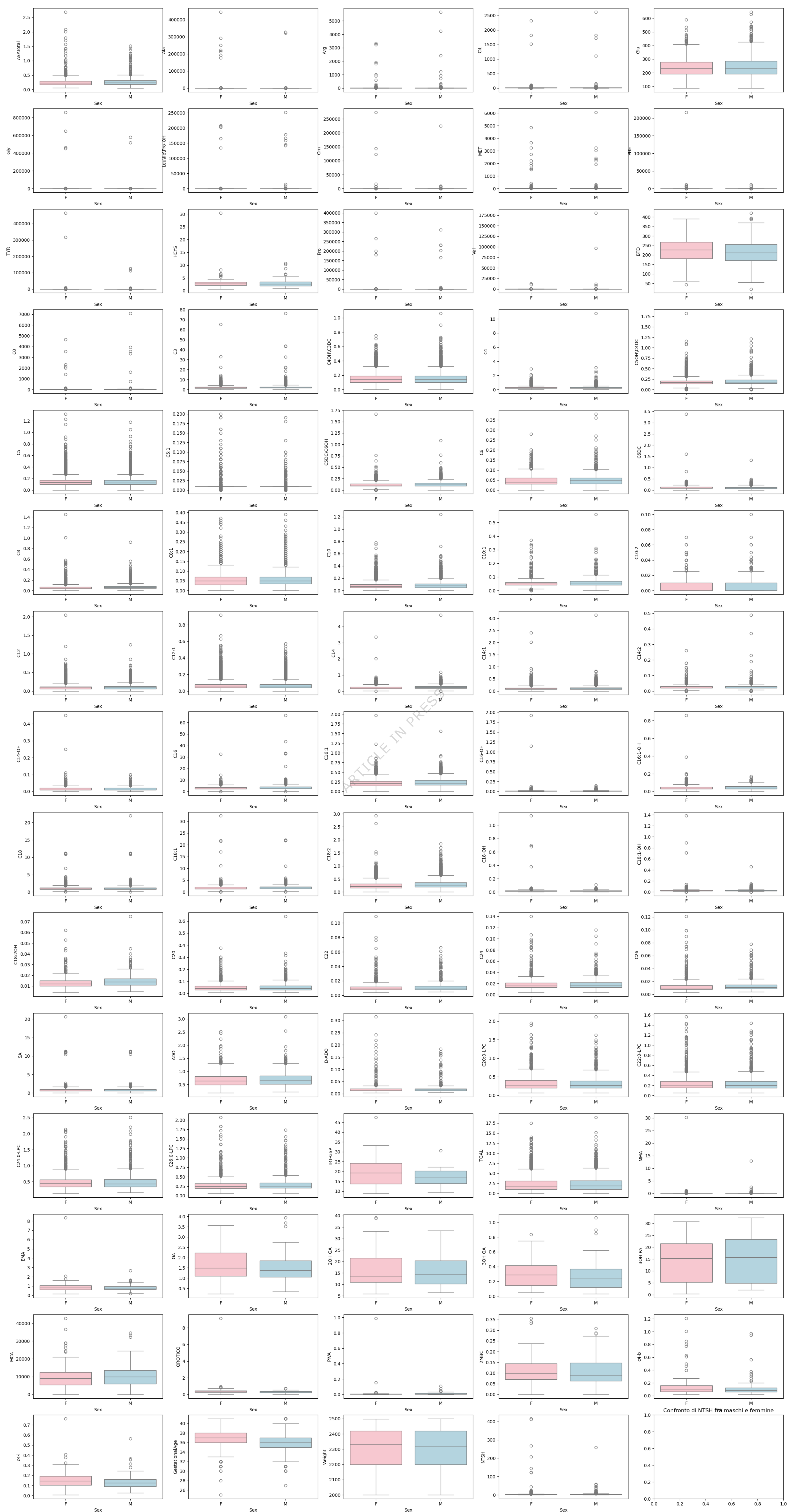
Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
Extended Lombardy's Neonatal Screening Dataset	Extended Lombardy's Neonatal Screening Dataset	10.5281/zenodo.16411149	https://doi.org/10.5281/zenodo.16411149	https://zenodo.org/records/16411149 https://zenodo.org/records/16411149/files/ExtendedLombardyNeonatalScreeningDataset.zip?download=1

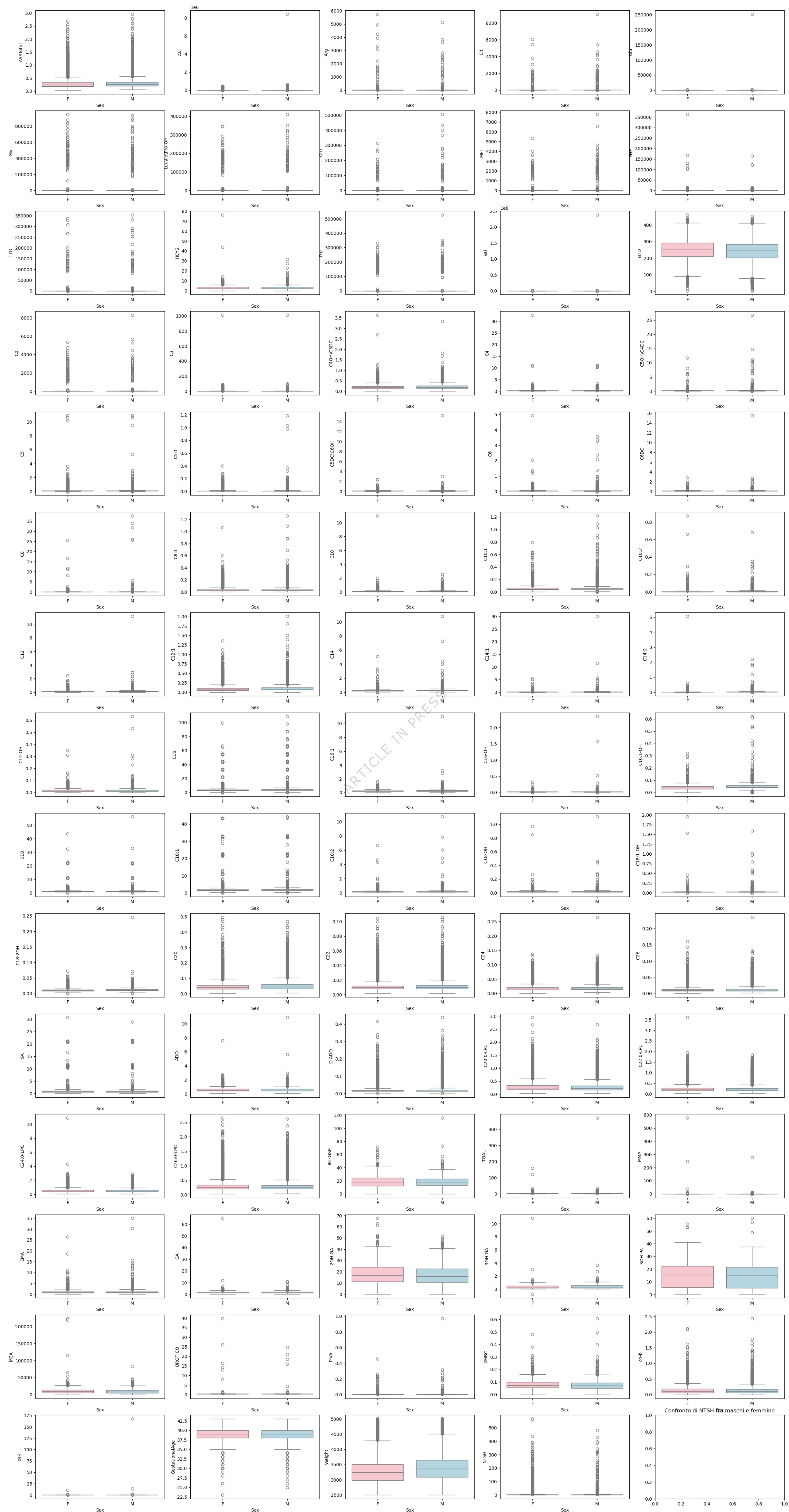












Extended Lombardy's Neonatal Screening Dataset

Lopiano, G. ^{1,*}, Alberti, L. ², Cereda, C. ², and Cabitza, F. ¹

¹University of Milano Bicocca, Computer Science, Milano, 20125, Italy

²Asst fatebenefratelli sacco, Milano, , Italy

*corresponding author: Gloria Lopiano (g.lopiano1@campus.unimib.it)

ABSTRACT

Neonatal Screening (NS) is a mandatory secondary prevention program established in the 1960s to detect rare genetic disorders in newborns. Early diagnosis through NS enables timely interventions, helping to prevent severe health complications. Despite its long-standing implementation, there remains a scarcity of publicly available datasets to support medical research and academic analysis.

In this work, we present a comprehensive dataset containing blood samples and associated information from nearly a decade of Neonatal Screening in Lombardy. This dataset offers valuable insights for medical studies on genetic disorders, early diagnostics, and treatment efficacy. Additionally, it can support academic research in data-driven healthcare analysis and policy development. By making this dataset publicly available, we aim to enhance research opportunities in neonatal health and contribute to the improvement of early intervention strategies.

Background & Summary

Neonatal metabolic screening is a cornerstone of preventive medicine across the world¹, enabling early detection of rare but severe metabolic disorders since the 1960s. These disorders, if left undiagnosed, can result in significant morbidity or mortality, underscoring the necessity of accurate and timely screening methods²³. The Neonatal Screening program in the Lombardy Region targets four groups of conditions: endocrine disorders (congenital hypothyroidism and congenital adrenal hyperplasia), cystic fibrosis, inherited metabolic disorders (NBS), and genetic neuromuscular diseases (such as spinal muscular atrophy). The screened diseases can be rare (affecting no more than one person in 2,000), congenital (present at birth), and often hereditary. In Italy, there are clear rules and indications to deal with the full process⁴, in particular, reference is made to the Ministry of Health Decree of 13 October 2016 (<https://www.trovanorme.salute.gov.it/norme/dettaglioAtto?id=56764#articoli>) that allows each region to work on it independently but in the same way. The process begins with collecting a few drops of blood from each baby in every hospitals of the region, usually between 48–72 hours after birth, in accordance with current national and international reference standards. Then the blood drops are stored on barcoded cards called Guthrie card⁵ (DBS), along with demographic and medical data submitted manually to the system. Samples and data are then sent to Buzzi Hospital in Milan for analysis using Genetic Screening Processor⁶ and Mass Spectrometry⁷. Further details about the Lombardy's Neonatal Screening analysis procedure will be added in the methods section. Poor-quality samples are flagged for recollection, labeled as BIS (if the sample is the second) or Controllo (if the sample is to check the previous one). The Mayo Clinic's neonatal screening software called Collaborative Labour Integrated Reports (CLIR)(<https://clir.mayo.edu/Home/About>)⁸ flags samples with a high probability of being positive to one or more diseases. Doctors review these cases and may request a second sample for further analysis, repeating the process and measuring different markers taking in consideration the diseases to check. Final diagnoses are based on weight-specific and disease specific cut-off values developed by doctors.

To our knowledge, no large public datasets on newborn screening currently exist. Having access to a large volume of real newborn screening data is crucial for facilitating information sharing between screening centers. This not only enables the comparison of biochemical marker concentration distributions across different populations but also allows for the evaluation of data collection quality and diagnostic assignment processes. In particular, thanks to the group of positive cases that have managed to be collected, it will be possible to test new automatic diagnosis assignment techniques. For this reason, the proposed dataset⁹ can be exploited in both the medical and academic fields. While the fact that the information it contains is real makes it possible to understand certain characteristics of the Lombardy population and the distribution of the concentration of particular biochemical markers, its size, the technical measurement errors it contains and the large number of missing values make it an excellent dataset with which to test different data cleaning techniques¹⁰. Nowadays many synthetic data are used for many health projects¹¹ because of privacy issue or data availability and surely and certainly using them is a good way to curb such

problems. However, depending on the generative model used to create them, one may encounter different technical difficulties and levels of realism¹².

The proposed dataset contains many observations from those that are usually fed into the model supporting the diagnostic process. Originally, the extracted observations contain samples of 10 years (from september 2012 to december 2022). The dataset originally comprised 985,792 rows and 266 features. The majority of the columns (235) correspond to biochemical marker concentration measurements, while 17 are binary features. The remaining columns contain information about the modules, such as the city of residence, the mother's ethnicity, and the time of sample collection. To ensure data quality and address potential errors, the dataset was processed, resulting in a reduced size.. As explained in the methodology section, this involved the removal of several variables and observations. Additionally, thanks to the contributions of physicians, diagnoses from the years 2016–2021 were integrated into the dataset. During the observed period, 198 positive cases were recorded. However, not every disease included in the screening had at least one positive case; rather, these 198 cases correspond to only 29 different diseases.

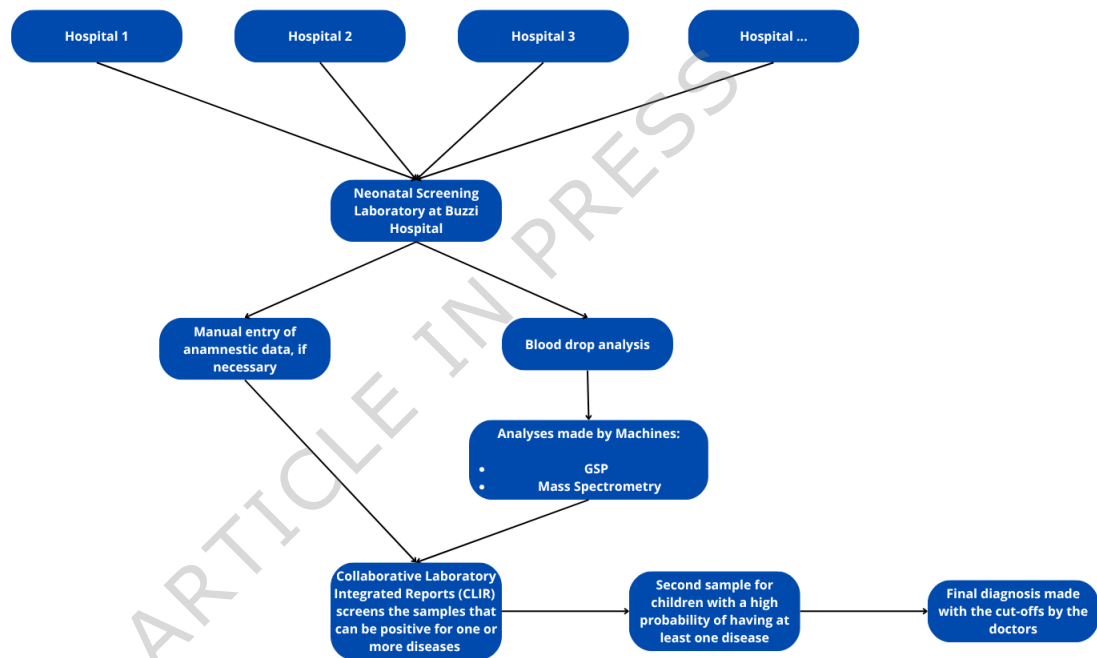


Figure 1. Flow Chart Process

Methods

The figure 1 illustrates the data-collection process organized by Lombardy region. Dried blood spot (DBS) samples and associated data collected at hospitals across Lombardy are sent to the Neonatal Screening Laboratory of Buzzi Hospital (Milan). This section details the steps for data collection, laboratory analysis, and data preparation/cleaning.

Ethical Consideration

Informed consent was obtained from all participating mothers through a signed document at the time of delivery. Pursuant to Italian legislation (e.g., Article 8, paragraph 2, Law No. 132/2025), and in accordance with local governance, the general informed consent permits secondary use of data stripped of direct identifiers. There is no evidence of harm associated with the DBS collection procedure. All procedures comply with applicable privacy regulations and institutional policies. The study was approved by the Territorial Ethics Committee 1 of Lombardy (CET 4–2023, 12 July 2023).

Collection

Between 48 and 72 hours after birth, a few drops of capillary blood are obtained from the newborn's heel (typically the right foot) and applied to a Guthrie card identified by a barcode. This time window balances neonatal physiological adaptation (reducing false-positive and false-negative results) with timely detection of target conditions such as congenital hypothyroidism and inborn errors of metabolism.

At the same time, a questionnaire records perinatal and maternal information. Anamnestic data are manually entered into the hospital information system by healthcare workers at each hospital, along with the associated barcode. Upon arrival at Buzzi Hospital, the barcode is scanned allowing the system to automatically retrieve the linked clinical/demographic information for analysis.

Shipping of Guthrie cards to Buzzi Hospital

In Lombardy, neonatal screening is centralized at Buzzi Hospital (Milan). Everyday hospitals send Guthrie cards in numbered parcels for processing. To ensure privacy, laboratory staff have defined two identifying variables that are excluded from the published dataset: one linked to the individual child and the other to the sample.

After associating the child's name with the barcode on the corresponding card, the system automatically integrates additional information. Occasional clerical errors may occur during manual handling (e.g., loss or physical damage of cards).

DBS analysis at the "V. Buzzi" Children Hospital

The DBSs were processed using the Neobase newborn screening kit (PerkinElmer, Milan, Italy). Single 3.2 mm spots were punched using a DBS puncher (PerkinElmer, Italy) into 96-well plates. A total of 100 μ L of the PerkinElmer extraction working solution was added to each well. The microplate was shaken for 45 min at 650 rpm at 45 °C and incubated for 120 min at room temperature to ensure the complete derivatization of the extracted succinylacetone. The plate was then quantified using a Waters TQD (Waters Corporation, Sesto San Giovanni, Italy) or Quattro Micro mass spectrometer in FIA-MS/MS. The concentrations were calculated by comparing the measured analyte intensities to those of the internal standards multiplied by the internal standard concentration and relative response factor. Eleven amino acids were analyzed: alanine (Ala), citrulline (Cit), arginine (Arg), glycine (Gly), leucine/isoleucine (Leu/Ile/Pro-OH), methionine (Met), ornithine (Orn), phenylalanine (Phe), tyrosine (Tyr), and valine (Val). Thirty-one acylcarnitines were analyzed: C0, C2, C3, C3DC, C4, C4DC, C5, C5:1, C5DC, C5OH, C6, C6DC, C8, C8:1, C10, C10:1, C10:2, C12, C12:1, C14, C14:1, C14:2, C14OH, C16, C16:1, C16OH, C16:1, C18, C18:1, C18:2, C18OH, and C18:1OH.¹³ The proposed dataset does not contain all of the cited biochemical entity. Calibration procedures are also taken into consideration for the analysis. In particular, each assay performed on the PerkinElmer GSP machine is calibrated using a specific calibration curve, while the mass spectrometer relies on internal standards for calibration.

Analysis of results and identification of positives:

Initial post-analytical evaluation is performed using the interactive web tool called Collaborative Laboratory Integrated Reports (CLIR)⁸. It is a very sophisticated software that has mainly three goals:

- Replacement of traditional cutoff values with continuous adjustments for age and other covariates of reference ranges shown as seamless percentile charts. CLIR reference ranges are derived by retrospective analysis of "big data", tens and even hundreds of thousands of data points from a growing worldwide community of collaborators.
- Creation of cumulative, covariate-adjusted disease ranges for all informative markers for target conditions, usually clustered by specialty and/or type of markers
- Post-analytical interpretive tools that integrate all relevant results into a single score. Tools are applicable either to the diagnosis and/or prognosis of a condition or to the differential diagnosis between pairs of conditions (for example benign variant vs. classic disease, responsive/not responsive to treatment).

The system automatically identifies babies with a high likelihood of having one or more target conditions. Laboratory physicians review the alerts and, when it is necessary, a repeat DBS is obtained for re-testing.

In the subsequent analysis, diagnosis is performed manually using conventional cut-off values, adjusted by four different weight categories (weight < 1500 g; 1500 g < weight < 2000 g; 2000 g < weight < 2500 g; weight > 2500 g) and three different sampling times.

Data preparation

Data extraction from the database was performed by the hospital's IT support team. Data preparation involved different scripts implemented in Python, primarily using the *polars* library.

Diagnostic information from 2016 to mid-2021 was integrated by merging a supplemental file via a unique identification code. The positive cases included 198 children diagnosed with 29 different metabolic diseases across various groups ¹.

Many samples contained multiple consecutive values within the same cell for several numerical variables. As the second value could not be reliably identified or interpreted, only the first value was retained, rounded to two decimal places. Composite variables derived from other fields were removed to avoid redundancy. Two additional biochemical markers, *s-TSH* and *s-I7OHP*, were excluded because they were measured in only 25 and 3 healthy newborns, respectively. Features judged unreliable, inconsistent or privacy-sensitive (*Allele 1*, *City*, *Ethnicity*, *Sampling*, *AnswerIX*, *SampleQuality*, *id*, and *SampleBarCode*) were removed. In particular, the removal of *id* and *SampleBarCode* ensured full anonymization of patient data.

All records were sorted by *SamTimeCollected*. Only the first observation for each patient ID was retained, except for affected children, for whom the diagnostic record was preserved, yielding 791,752 records.

After the feature reductions, the number of variables was 100. Consequently, data type were standardized. Since all binary variables indicate the presence or absence of a specific characteristic or condition, the high number of missing values resulted from many operators omitting negative responses, leaving the field empty. For this reason, missing values in binary variables were recoded as 0.

Handling of missing values is particularly critical, as "Ignoring or inadequately handling missing values can lead to biases and loss of statistical power"¹⁴. Despite this, we opted not to modify the large number of missing values in the dataset, even though proper handling is generally preferable to ignoring them¹⁵. This choice allows users to evaluate alternative strategies^{16–20}.

Two potentially informative variables *SamTimeReceived* and *SamTimeCollected*, were frequently incomplete or incorrectly recorded. These variables would have been particularly valuable if timestamps were available for all observations, but they were missing for most observations. Consequently, only the calendar date was retained and variable names were updated accordingly.

We then computed the interval between *SamTimeCollected* and *DateOfBirth* and excluded all observations with this value lower than 1 or greater than 3 days. After this filtering step, the final number of rows was reduced to 748,716.

Outliers were retained to preserve the dataset's Although several variables exhibited extreme values (including maxima exceeding three times the 99th percentile), the definition of outliers is context-dependent. We chose to retain all values as extracted from the hospital server in order to allow users to determine their own outlier definitions and apply appropriate correction methods as needed.

Data Records

The data are available on Zenodo <https://doi.org/10.5281/zenodo.16411149>⁹. The dataset is stored as a separate comma-separated values (CSV) file, where each record corresponds to the first sample collected from a newborn in Lombardy. It includes numeric, string, date, and binary columns.

The dataset consists of 748,716 rows and 100 columns. A complete list of feature names is provided in Table 2.

Technical Validation

Developing datasets is a process that requires significant responsibility and attention. Data quality is of fundamental importance, as it has a major impact on the reliability of extracted information. It is essential to ensure that this medical dataset conforms to the recommendations outlined by the IEEE in the paper *IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence*²¹.

This section outlines several procedures for verifying and inspecting data quality. The code file available in the repository includes checks to ensure the correctness of certain information. In particular:

- The fact that all babies with the variable *Premature* equal to 1 have a *GestationalAge* lower than 36 was inspected and eventually corrected.
- To select the first sample for each baby, the dataset was sorted taking in consideration the variable *SamTimeCollected*
- A check was made on the time window of sample acquisition by calculating the difference between date of collection and date of birth. Some observations were found to be far outside the 48-72 hour window of the health standard. In some cases the error is probably a typing error, while in others there may have been an unforeseen event that caused a delay. These samples, in case they did not belong to positive children, were removed from the published dataset.
- Other features that didn't deal with their original description were deleted.

There are also some medical shrewdness applied, in particular during the extraction of the biochemical markers concentrations like the periodic change of reagents and the daily sanitization of the spindle of PelkinElmer machine.

Data Availability

The dataset is available on Zenodo (<https://doi.org/10.5281/zenodo.16411149>)⁹ while the Supplementary Information file is available on Github (https://github.com/GloLo/NSE_Lombardy.git) in the Supplementary Information folder.

Code availability

The code was written in Python language (v. 3.12.4) using the polars (v. 1.8.2), numpy (v. 1.26.4), packages. The generation and validation code is available on GitHub at the following URL: https://github.com/GloLo/NSE_Lombardy.git

Acknowledgements

The authors thank the Neonatal Screening of Lombardy for the provision of data and Luca Marconi for his effort.

Author contributions statement

G.L.: Responsible for cleaning the data, writing the paper and the figures and tables in it

L.A.: Medical support for clarification of values of variables in the dataset and their meaning.

C.C.: Responsible for data collection and extraction. Head of the Neonatal Screening Laboratory of Lombardy

F.C.: Supervisor of data cleaning choices and reviewer of the article.

Competing interests

The authors declare no competing interests. The authors acknowledge funding support provided by PNRR funds (CUP: H41J24000400002) and Fastweb SpA.

Figures & Tables

Fig. 1 Flow chart of the data collection process

Fig. 2 Number of missing values in the numerical features

Fig. 3 Babies with a weight lower than 1500g

Fig. 4 Babies with a weight between 2000g and 1500g

Fig. 5 Babies with a weight between 2500g and 2000g

Fig. 6 Babies with a weight higher than 2500g

Tab. 1 Summary of the diseases of neonatal screening program of Lombardy.

Tab. 2 Summary of dataset's features

Tab.3 Babies with a weight lower than 1500g

Tab.4 Babies with a weight between 2000g and 1500g

Tab.5 Babies with a weight between 2500g and 2000g

Tab. 6 Babies with a weight higher than 2500g

Tab. 7 Number of positive children per weight group



Figure 2. Number of missing values



Figure 3. Comparison between emale and Male < 1500g



Figure 4. Comparison between Female and Male 1500-2000 g

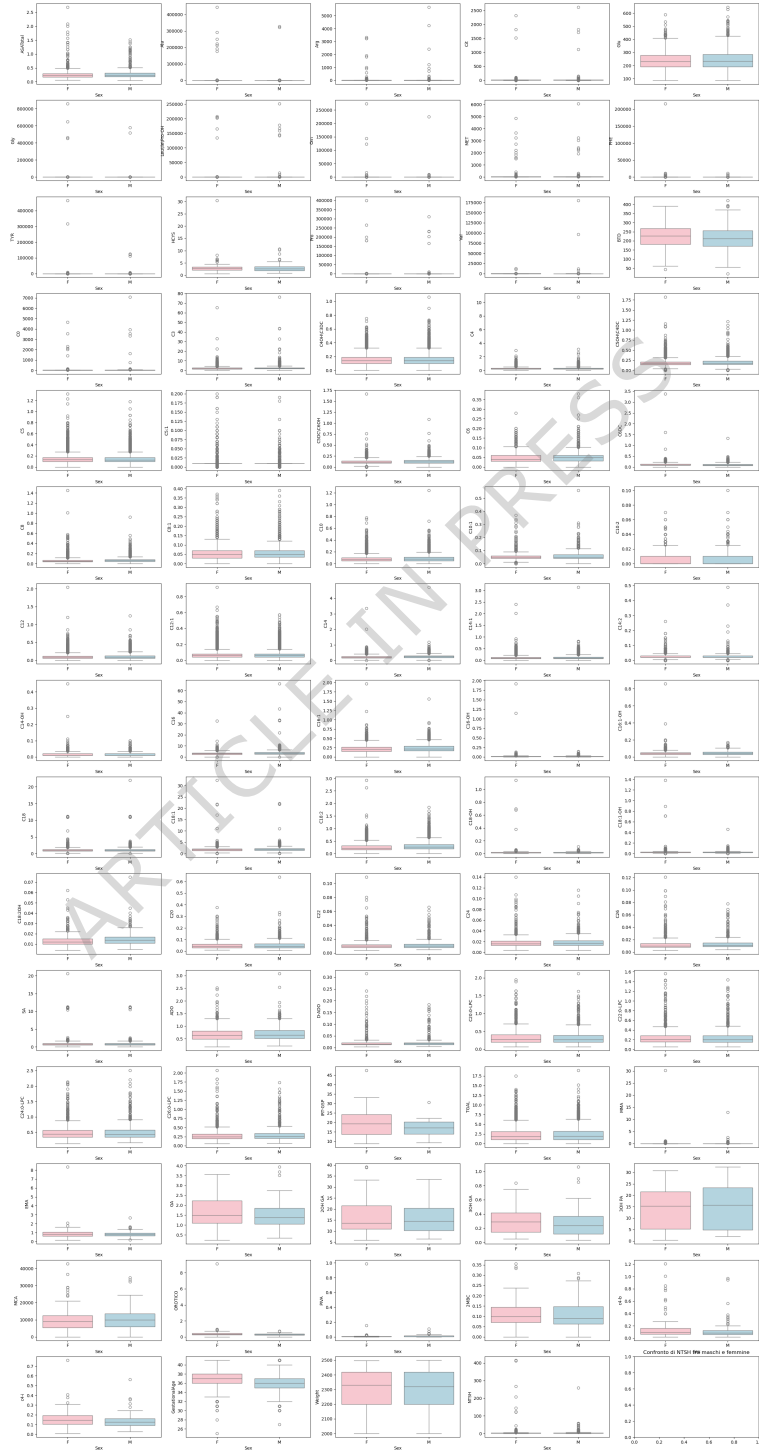


Figure 5. Comparison between Female and Male 2000-2500 g



Figure 6. Comparison between Female and Male >2500 g

Acronym	Name of the disease	Category
ASA	argininosuccinic acidemia	UCD (urea cycle disorders)
IBG	acyl-CoA dehydrogenase deficiency	AO (organic acidemias)
CIT1	citrullinemia type I	UCD (urea cycle disorders)
CBS	Homocystinuria	AA (aminoacidopathies)
HPA	benign hyperphenylalanine	AA (aminoacidopathies)
PA	propionic acidemia	AO (organic acidemias)
CUD	carnitine transport deficiency	FAO (fatty acid oxidation disorders)
TYR3	tyrosinaemia type III	AA (aminoacidopathies)
CPT2	carnitine palmitoyltransferase II deficiency	FAO (fatty acid oxidation disorders)
VLCAD	very long-chain acyl-CoA dehydrogenase deficiency	FAO (fatty acid oxidation disorders)
MMA	generic hereditary metabolic disease	AO (organic acidemias)
IVA	isovaleric acidemia	AO (organic acidemias)
PHE	phenylketonuria	AA (aminoacidopathies)
MTHFR	homocystinuria (severe MTHFR defect)	AA (aminoacidopathies)
MCAD	medium-chain acyl-CoA dehydrogenase deficiency	FAO (fatty acid oxidation disorders)
BKT	beta ketothiolase deficiency	AO (organic acidemias)
SCAD	short-chain acyl-CoA dehydrogenase deficiency	FAO (fatty acid oxidation disorders)
CblD	methylmalonic acidemia with homocystinuria (Cbl D deficiency)	AO (organic acidemias)
GA2	glutaric acidemia type II	FAO (fatty acid oxidation disorders)
GA1	glutaric acidemia type I	AO (organic acidemias)
ARG	arginemia	UCD (urea cycle disorders)
CPS1	Carbamoyl phosphate synthetase 1 deficiency (CPSID)	AA (aminoacidopathies)
2MBG	2-methylbutyryl-CoA dehydrogenase deficiency	AO (organic acidemias)
PKU	phenylketonuria	AA (aminoacidopathies)
CblC	methylmalonic acidemia with homocystinuria (Cbl C deficiency)	AO (organic acidemias)
ARG1	Arginase-1 Deficiency	UCD (urea cycle disorders)
CblA	methylmalonic acidemia (Cbl A deficiency)	AO (organic acidemias)
TFP EXITUS	mitochondrial trifunctional protein deficiency	FAO (fatty acid oxidation disorders)
3MCC	3-methylcrotonyl-CoA carboxylase deficiency	AO (organic acidemias)
BIOPT (BS)	Biopterin cofactor biosynthesis deficiency	AA (aminoacidopathies)
BIOPT (REG)	Biopterin cofactor regeneration deficiency	AA (aminoacidopathies)
TYR I	Tyrosinemia type I	AA (aminoacidopathies)
TYR II	Tyrosinemia type II	AA (aminoacidopathies)
MSUD	Maple syrup urine disease	AA (aminoacidopathies)
HCY	Homocystinuria (CBS deficiency)	AA (aminoacidopathies)
HMG	3-Hydroxy-3-methylglutaric acidemia	OA (organic acidemias)
MUT	Methylmalonic acidemia (Mut type)	OA (organic acidemias)
CblB	Methylmalonic acidemia (Cbl-B type)	OA (organic acidemias)
CblC	Methylmalonic acidemia (Cbl-C type)	OA (organic acidemias)
CblD	Methylmalonic acidemia (Cbl-D type)	OA (organic acidemias)
2MBG	2-Methylbutyryl-CoA dehydrogenase deficiency	OA (organic acidemias)
MAL	Malonic aciduria	OA (organic acidemias)
MCD	Multiple carboxylase deficiency	OA (organic acidemias)
CIT II	Citrullinemia type II	UCD (urea cycle disorders)
CPT Ia	Carnitine palmitoyltransferase I deficiency	FAO (fatty acid oxidation disorders)
CACT	Carnitine-acylcarnitine translocase deficiency	FAO (fatty acid oxidation disorders)
TFP	Mitochondrial trifunctional protein deficiency	FAO (fatty acid oxidation disorders)
LCHAD	Long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency	FAO (fatty acid oxidation disorders)

Table 1. Summary of the diseases included in the neonatal screening program of the Lombardy region. A complete version of the table is provided in the Supplementary Information. The table shown here presents a subset of the full list of diseases screened in the SNE. The complete version is available in the Supplementary Information.

Feature	Description	Percentage of Missing Values
ASATotal	argininosuccinic acid	93.33%
Ala	alanine	0.02%
Arg	arginine	0.02%
Cit	citrulline	0.02%
Glu	glutamic acid	93.33%
Gly	glycine	0.02%
Leu/Ile/Pro-OH	leucine/ isoleucine/hydroxyproline	0.02%
Orn	ornithine	0.02%
MET	methionine	0.02%
PHE	phenylalanine	0.02%
TYR	tyrosine	0.02%
HCYS	homocysteine	99.63%
Pro	proline	0.02%
Val	valine	0.02%
BTD	humanbiotinidase activity	68.74%
C0	free carnitine	0.02%
C3	propionylcarnitine	0.02%
C4OH/C3DC	malonylcarnitine / 3-hydroxybutyrylcarnitine	0.02%
C4	butyrylcarnitine	0.02%
C5OH/C4DC	methylmalonylcarnitine / 3-hydroxyisovalerylcarnitine	0.02%
C5	isovalerylcarnitine	0.02%
C5:1	tiglylcarnitine	0.02%
C5DC/C6OH	glutaryl carnitine / 3-hydroxyhexanoylcarnitine	0.02%
C6	hexanoylcarnitine	0.02%
C6DC	adipoylcarnitine	0.02%
C8	octanoylcarnitine	0.02%
C8:1	octenoylcarnitine	0.02%
C10	decanoylcarnitine	0.02%
C10:1	decenoylcarnitine	0.02%
C10:2	decadienoylcarnitine	0.02%
C12	dodecanoylcarnitine	0.02%
C12:1	dodecenoylcarnitine	0.02%
C14	tetradecanoylcarnitine (myristoylcarnitine)	0.02%
C14:1	tetradecenoylcarnitine	0.02%
C14:2	tetradecadienoylcarnitine	0.02%
C14-OH	3-hydroxy tetradecanoylcarnitine	0.02%
C16	hexadecanoylcarnitine (palmitoylcarnitine)	0.02%
C16:1	hexadecenoylcarnitine	0.02%
C16-OH	3-hydroxy-hexadecanoylcarnitine	0.02%
C16:1-OH	3-hydroxy-hexadecenoylcarnitine	0.02%
C18	octadecanoylcarnitine (stearoylcarnitine)	0.02%
C18:1	octadecenoylcarnitine (oleoylcarnitine)	0.02%
C18:2	octadecadienoylcarnitine (linoleoylcarnitine)	0.02%
C18-OH	3-hydroxy-octadecanoylcarnitine	0.02%
C18:1-OH	3-hydroxy-octadecenoylcarnitine	0.02%
C18:2OH	3-hydroxy-octadecadienoylcarnitine	93.33%
C20	eicosanoylcarnitine (arachidoylcarnitine)	93.33%
C22	docosanoylcarnitine (Behenoylcarnitine)	93.33%
C24	tetracosanoylcarnitine (lignoceroylcarnitine)	93.33%
C26	hexacosanoylcarnitine (cerotoylcarnitine)	93.33%
SA	succinylacetone	0.02%
ADO	adenosine	93.33%
D-ADO	2-deoxyadenosine	93.33%

Table 2. Summary of the features of the dataset. The table shown here presents a subset of the feature kept in the dataset. The complete version is available in the Supplementary Information.

Name	null_count	mean	std	min	25%	50%	75%	max	99%	1%	Kurtosis
ASATotal	5603.0	0.373	0.213	0.092	0.247	0.321	0.448	2.514	1.12	0.124	29.796
Ala	0.0	730.557	12666.073	0.0	189.26	248.6	337.76	519538.705	818.36	91.78	975.504
Arg	0.0	117.452	6388.356	0.0	13.14	24.37	40.7	493407.175	139.26	2.45	5957.934
Cit	0.0	19.816	84.738	0.0	11.84	15.63	20.11	3118.475	45.44	4.76	999.406
Glu	5603.0	250.195	82.511	96.167	188.029	240.361	299.02	559.736	501.395	114.199	0.664
Gly	0.0	1240.835	22900.921	0.0	350.65	443.192	554.67	870900.195	1016.38	163.54	958.082
Leu/Ile/Pro-OH	0.0	632.043	10637.838	41.43	185.099	228.58	278.16	349314.915	606.26	74.7	790.401
Orn	0.0	354.193	7522.102	6.37	101.49	138.59	184.42	401343.345	363.71	35.11	1899.411
MET	0.0	68.362	1666.985	3.37	22.07	30.82	42.8	127393.69	188.13	8.8	5704.24
PHE	0.0	211.537	3995.467	19.93	66.23	78.9	94.38	176230.335	205.16	40.32	1183.296
TYR	0.0	309.076	10502.069	0.0	49.35	75.31	117.778	625571.97	759.32	8.25	3081.437
HCYS	5929.0	2.117	1.254	0.463	1.233	1.776	2.743	6.506	6.506	0.463	2.55
Pro	0.0	1556.58	72720.957	27.55	138.48	178.73	235.57	5541527.355	561.14	65.02	5645.226
Val	0.0	930.751	18939.656	30.29	159.12	202.84	251.33	808815.97	462.67	62.12	1247.524
BTDC	4187.0	217.695	67.249	36.681	164.232	213.402	270.806	423.957	356.701	83.775	-0.787
C0	0.0	39.748	112.202	6.18	21.97	31.24	47.56	6869.475	110.23	10.1	2690.049
C3	0.0	3.034	1.628	0.31	2.002	2.75	3.75	54.655	7.73	0.81	175.934
C4OH/C3DC	0.0	0.144	0.138	0.0	0.1	0.13	0.18	9.41	0.36	0.04	3449.414
C4	0.0	0.454	0.279	0.06	0.29	0.39	0.54	8.1	1.423	0.14	106.621
C5OH/C4DC	0.0	0.19	0.077	0.0	0.14	0.18	0.22	1.37	0.43	0.07	28.895
C5	0.0	0.384	0.208	0.05	0.25	0.34	0.46	2.62	1.149	0.1	9.587
C5:1	0.0	0.015	0.008	0.0	0.01	0.01	0.02	0.11	0.04	0.0	7.426
C5DC/C6OH	0.0	0.125	0.11	0.0	0.09	0.11	0.15	7.36	0.29	0.044	3177.522
C6	0.0	0.057	0.036	0.0	0.04	0.05	0.07	1.29	0.17	0.02	267.467
C6DC	0.0	0.077	0.037	0.0	0.05	0.07	0.1	0.52	0.19	0.01	6.461
C8	0.0	0.091	0.06	0.0	0.059	0.08	0.11	1.2	0.31	0.02	44.849
C8:1	0.0	0.094	0.063	0.0	0.05	0.08	0.12	0.59	0.32	0.017	6.813
C10	0.0	0.059	0.048	0.0	0.04	0.05	0.07	1.66	0.2	0.01	321.287
C10:1	0.0	0.063	0.048	0.0	0.03	0.05	0.07	0.69	0.25	0.01	21.253
C10:2	0.0	0.01	0.012	0.0	0.0	0.01	0.015	0.17	0.05	0.0	12.355
C12	0.0	0.041	0.023	0.0	0.03	0.04	0.05	0.39	0.12	0.01	40.089
C12:1	0.0	0.032	0.021	0.0	0.02	0.03	0.04	0.39	0.1	0.01	58.545
C14	0.0	0.152	0.072	0.01	0.1	0.14	0.19	0.65	0.39	0.05	3.422
C14:1	0.0	0.072	0.04	0.0	0.05	0.062	0.09	0.82	0.19	0.02	43.79
C14:2	0.0	0.03	0.018	0.0	0.02	0.03	0.04	0.34	0.08	0.01	39.224
C14-OH	0.0	0.008	0.005	0.0	0.007	0.01	0.01	0.09	0.02	0.0	33.626
C16	0.0	1.695	0.991	0.08	1.08	1.51	2.07	22.255	4.54	0.45	95.467
C16:1	0.0	0.134	0.069	0.01	0.09	0.12	0.16	0.63	0.38	0.04	5.428
C16-OH	0.0	0.012	0.009	0.0	0.01	0.01	0.01	0.46	0.03	0.0	1312.89
C16:1-OH	0.0	0.039	0.015	0.0	0.03	0.04	0.05	0.21	0.083	0.01	5.49
C18	0.0	0.91	0.526	0.14	0.65	0.85	1.09	21.822	1.98	0.31	527.893
C18:1	0.0	1.506	0.723	0.04	1.06	1.39	1.82	21.4	3.45	0.49	124.232
C18:2	0.0	0.631	0.412	0.01	0.38	0.57	0.8	11.135	1.68	0.1	208.88
C18-OH	0.0	0.007	0.009	0.0	0.0	0.01	0.01	0.47	0.02	0.0	1568.854
C18:1-OH	0.0	0.021	0.016	0.0	0.01	0.02	0.02	0.72	0.05	0.01	1221.963
C18:2OH	5603.0	0.023	0.009	0.004	0.016	0.021	0.029	0.058	0.052	0.008	1.036
C20	5603.0	0.034	0.016	0.007	0.024	0.03	0.039	0.139	0.092	0.014	11.368
C22	5603.0	0.012	0.007	0.004	0.008	0.01	0.013	0.092	0.036	0.005	57.17
C24	5603.0	0.014	0.008	0.004	0.009	0.012	0.016	0.089	0.039	0.005	30.046
C26	5603.0	0.013	0.008	0.004	0.008	0.011	0.015	0.118	0.035	0.005	78.844
SA	0.0	0.831	0.533	0.0	0.59	0.81	1.03	11.6	1.65	0.15	220.395
ADO	5603.0	1.111	0.505	0.234	0.81	1.027	1.306	4.605	2.724	0.409	12.292
D-ADO	5603.0	0.022	0.02	0.007	0.014	0.018	0.024	0.313	0.078	0.008	135.48
C20:0-LPC	5603.0	0.369	0.235	0.037	0.219	0.31	0.457	1.772	1.271	0.078	7.915
C22:0-LPC	5603.0	0.364	0.221	0.042	0.22	0.307	0.439	1.865	1.085	0.082	7.411
C24:0-LPC	5603.0	0.477	0.232	0.07	0.327	0.432	0.566	1.875	1.236	0.161	6.381
C26:0-LPC	5603.0	0.33	0.211	0.045	0.194	0.281	0.416	2.216	0.988	0.09	21.51
IRT-GSP	5959.0	26.718	7.849	16.436	25.828	26.081	33.565	38.768	38.768	16.436	-0.989
TGAL	4189.0	1.788	1.571	0.0	0.688	1.311	2.461	11.306	7.06	0.0	3.701
MMA	310.0	0.006	0.205	0.0	0.0	0.0	0.0	15.078	0.0	0.0	5187.711
EMA	5929.0	0.933	0.486	0.216	0.6	0.82	1.162	2.349	2.349	0.216	0.543
GA	5929.0	1.938	0.914	0.696	1.268	1.637	2.281	4.378	4.378	0.696	0.513
2OH GA	5929.0	15.256	7.673	3.218	9.298	14.638	18.956	36.764	36.764	3.218	0.192
3OH GA	5929.0	0.348	0.2	0.012	0.228	0.369	0.479	0.744	0.744	0.012	-0.742
3OH PA	5929.0	17.124	8.591	2.807	7.838	17.86	24.194	36.022	36.022	2.807	-0.945
MCA	5929.0	12270.046	8051.695	0.194	6599.051	12917.59	16230.858	37212.579	37212.579	0.194	0.922
OROTICO	5929.0	0.47	0.277	0.06	0.285	0.375	0.683	1.076	1.076	0.06	-0.631
PIVA	5929.0	0.019	0.014	0.0	0.006	0.018	0.027	0.062	0.062	0.0	0.806
2MBC	5929.0	0.293	0.158	0.062	0.172	0.264	0.365	0.751	0.751	0.062	0.488
c4-b	5929.0	0.316	0.471	0.031	0.073	0.135	0.301	2.28	2.28	0.031	7.146
c4-i	5929.0	1.196	5.732	0.052	0.161	0.244	0.38	35.115	35.115	0.052	31.998
GestationalAge	0.0	29.775	2.905	23.0	28.0	30.0	32.0	41.0	36.0	23.0	-0.012
Weight	0.0	1154.136	280.631	350.0	956.0	1210.0	1400.0	1499.0	1495.0	460.0	-0.579
NTSH	0.0	2.055	4.8	0.0	0.916	1.571	2.54	348.0	8.15	0.186	4520.995

Table 3. Statistics of the observations with a weight lower than 1500 g (5966 babies)

Name	null_count	mean	std	min	25%	50%	75%	max	99%	1%	Kurtosis
ASATotal	8635.0	0.279	0.207	0.08	0.176	0.233	0.306	1.885	1.271	0.09	27.761
Ala	4.0	353.673	5363.066	0.0	179.4	231.74	300.06	384289.65	596.48	89.23	3765.2
Arg	4.0	17.533	81.008	0.0	5.21	9.78	18.68	5661.67	82.89	1.01	2811.18
Cit	4.0	14.368	31.046	0.0	9.92	12.84	16.69	2023.445	33.43	3.94	3827.177
Glu	8635.0	242.547	75.106	95.927	191.065	232.338	283.528	608.563	497.136	108.938	1.797
Gly	4.0	854.31	14959.482	0.0	318.4	401.64	502.4	687520.44	965.51	141.12	1381.13
Leu/Ile/Pro-OH	4.0	218.1	3843.236	0.0	114.51	148.176	196.47	346368.88	380.84	62.18	7294.351
Orn	4.0	166.65	4068.511	0.0	68.03	96.23	137.11	354289.59	315.41	26.85	6532.994
MET	4.0	27.403	99.329	0.0	16.89	22.47	29.78	6664.185	78.57	7.97	3863.896
PHE	4.0	97.235	2015.987	0.0	49.5	59.96	73.75	140135.975	143.92	30.32	4543.33
TYR	4.0	1486.149	132230.53	0.0	61.37	86.55	125.55	12661131.27	373.27	21.67	9162.989
HCYS	9126.0	2.16	1.218	0.701	1.359	1.74	2.776	6.982	6.982	0.701	3.848
Pro	4.0	305.776	7904.06	0.0	121.07	151.84	190.4	655701.47	357.79	62.42	5360.549
Val	4.0	156.508	2010.787	0.0	86.17	115.55	162.32	192185.985	334.292	44.15	9071.657
BTDC	6479.0	212.682	61.366	4.168	165.722	208.395	259.334	375.548	345.192	95.348	-0.673
C0	4.0	29.935	70.369	0.0	18.5	24.82	34.21	4040.0	79.94	9.83	2260.065
C3	4.0	2.587	1.814	0.0	1.66	2.28	3.14	76.27	6.92	0.74	634.76
C4OH/C3DC	4.0	0.114	0.054	0.0	0.08	0.1	0.14	0.64	0.3	0.03	4.077
C4	4.0	0.302	0.207	0.0	0.19	0.26	0.35	10.86	0.96	0.09	768.165
C5OH/C4DC	4.0	0.175	0.075	0.0	0.13	0.17	0.21	3.73	0.37	0.07	570.317
C5	4.0	0.211	0.118	0.0	0.13	0.18	0.25	1.39	0.65	0.07	7.564
C5:1	4.0	0.01	0.007	0.0	0.01	0.01	0.01	0.21	0.03	0.0	95.995
C5DC/C6OH	4.0	0.113	0.117	0.0	0.08	0.1	0.13	7.62	0.25	0.04	3227.287
C6	4.0	0.047	0.022	0.0	0.03	0.04	0.06	0.5	0.12	0.01	29.319
C6DC	4.0	0.082	0.039	0.0	0.06	0.08	0.1	0.36	0.2	0.013	2.082
C8	4.0	0.058	0.049	0.0	0.04	0.05	0.07	3.49	0.16	0.01	2596.313
C8:1	4.0	0.065	0.037	0.0	0.04	0.06	0.08	0.53	0.19	0.01	7.599
C10	4.0	0.058	0.041	0.0	0.036	0.05	0.07	2.22	0.18	0.01	837.104
C10:1	4.0	0.049	0.027	0.0	0.03	0.04	0.06	0.42	0.14	0.01	16.369
C10:2	4.0	0.005	0.007	0.0	0.0	0.0	0.01	0.11	0.03	0.0	9.069
C12	4.0	0.062	0.036	0.0	0.04	0.05	0.08	0.47	0.19	0.02	11.566
C12:1	4.0	0.041	0.028	0.0	0.02	0.03	0.05	0.42	0.14	0.01	18.883
C14	4.0	0.198	0.081	0.0	0.14	0.188	0.24	0.77	0.44	0.06	2.37
C14:1	4.0	0.081	0.043	0.0	0.05	0.07	0.1	0.7	0.22	0.02	14.237
C14:2	4.0	0.021	0.01	0.0	0.01	0.02	0.03	0.1	0.05	0.01	3.772
C14-OH	4.0	0.009	0.006	0.0	0.01	0.01	0.01	0.16	0.03	0.0	68.921
C16	4.0	2.546	1.178	0.0	1.76	2.38	3.12	43.95	5.96	0.79	174.14
C16:1	4.0	0.182	0.086	0.0	0.12	0.17	0.23	0.81	0.44	0.05	1.44
C16-OH	4.0	0.014	0.007	0.0	0.01	0.01	0.02	0.13	0.036	0.0	12.067
C16:1-OH	4.0	0.042	0.016	0.0	0.03	0.04	0.05	0.3	0.09	0.02	15.087
C18	4.0	0.964	0.433	0.0	0.71	0.9	1.14	11.16	2.04	0.39	189.995
C18:1	4.0	1.726	0.713	0.0	1.28	1.64	2.07	32.474	3.51	0.65	389.261
C18:2	4.0	0.379	0.264	0.0	0.228	0.33	0.471	10.81	1.11	0.09	521.702
C18-OH	4.0	0.009	0.005	0.0	0.01	0.01	0.01	0.05	0.02	0.0	2.719
C18:1-OH	4.0	0.022	0.01	0.0	0.02	0.02	0.03	0.136	0.05	0.01	5.932
C18:2OH	8635.0	0.017	0.006	0.002	0.013	0.016	0.02	0.054	0.035	0.006	3.392
C20	8635.0	0.035	0.023	0.007	0.022	0.03	0.041	0.29	0.121	0.012	34.45
C22	8635.0	0.011	0.007	0.003	0.008	0.009	0.012	0.082	0.045	0.005	35.211
C24	8635.0	0.014	0.008	0.004	0.009	0.012	0.017	0.077	0.045	0.005	16.589
C26	8635.0	0.013	0.009	0.003	0.008	0.011	0.014	0.106	0.059	0.005	40.812
SA	4.0	0.804	0.383	0.0	0.57	0.78	1.02	11.365	1.62	0.11	152.047
ADO	8635.0	0.832	0.326	0.25	0.608	0.785	0.988	2.488	1.955	0.29	3.361
D-ADO	8635.0	0.02	0.02	0.006	0.013	0.017	0.021	0.244	0.13	0.008	67.734
C20:0-LPC	8635.0	0.363	0.248	0.057	0.224	0.299	0.418	2.443	1.522	0.096	19.089
C22:0-LPC	8635.0	0.293	0.216	0.054	0.175	0.239	0.342	2.013	1.418	0.085	17.623
C24:0-LPC	8635.0	0.484	0.265	0.167	0.336	0.416	0.542	2.966	1.511	0.19	18.88
C26:0-LPC	8635.0	0.31	0.209	0.064	0.192	0.255	0.351	1.976	1.228	0.093	17.867
IRT-GSP	9160.0	20.015	8.662	9.177	13.733	18.386	27.495	31.992	31.992	9.177	-1.507
TGAL	6480.0	2.197	2.033	0.0	0.915	1.723	2.981	57.882	7.896	0.0	209.056
MMA	467.0	0.003	0.044	0.0	0.0	0.0	0.0	1.933	0.0	0.0	768.194
EMA	9126.0	1.095	1.352	0.362	0.685	0.872	1.241	9.74	9.74	0.362	34.988
GA	9126.0	2.065	1.364	0.276	1.131	1.619	2.776	6.467	6.467	0.276	1.557
2OH GA	9126.0	18.476	11.234	4.667	11.094	16.275	21.322	57.634	57.634	4.667	2.343
3OH GA	9126.0	0.289	0.21	0.014	0.122	0.249	0.467	0.956	0.956	0.014	0.552
3OH PA	9126.0	15.802	9.13	1.27	6.716	15.941	23.242	34.826	34.826	1.27	-0.987
MCA	9126.0	10752.253	6605.581	0.0	6248.572	9864.037	14863.84	27489.156	27489.156	0.0	-0.319
OROTICO	9126.0	0.402	0.219	0.08	0.25	0.326	0.553	0.958	0.958	0.08	-0.352
PIVA	9126.0	0.009	0.009	0.0	0.001	0.008	0.016	0.031	0.031	0.0	-0.27
2MBC	9126.0	0.184	0.124	0.0	0.1	0.158	0.257	0.538	0.538	0.0	0.205
c4-b	9126.0	0.28	0.321	0.039	0.087	0.133	0.36	1.238	1.238	0.039	1.782
c4-i	9126.0	0.203	0.121	0.04	0.105	0.183	0.236	0.515	0.515	0.04	0.23
GestationalAge	0.0	33.912	1.99	23.0	32.0	34.0	35.0	41.0	38.0	30.0	-0.136
Weight	0.0	1805.237	132.193	1500.0	1700.0	1825.0	1920.0	1999.0	1993.0	1520.0	-0.942
NTSH	0.0	2.322	3.501	0.0	1.117	1.803	2.898	290.0	8.68	0.304	4967.327

Table 4. Statistics of the observations with a weight between 1500 g and 2000 g (9172 babies)

Name	null_count	mean	std	min	25%	50%	75%	max	99%	1%	Kurtosis
ASATotal	33231.0	0.269	0.179	0.05	0.176	0.232	0.305	2.692	1.115	0.091	36.733
Ala	6.0	325.408	4502.878	0.0	194.69	242.68	302.13	445301.98	541.324	107.75	5443.927
Arg	6.0	10.914	53.557	0.0	4.49	7.8	12.86	5661.43	43.15	0.99	5927.453
Cit	6.0	13.416	27.059	0.0	9.698	12.34	15.61	2621.565	28.7	3.89	5805.841
Glu	33231.0	243.411	72.803	87.504	191.468	232.42	282.835	646.134	466.476	122.221	1.935
Gly	6.0	535.763	7837.197	0.0	337.32	417.65	511.66	858862.565	914.03	157.97	7403.405
Leu/Ile/Pro-OH	6.0	195.043	3185.249	0.0	111.23	132.74	159.15	251280.165	277.99	70.09	3713.23
Orn	6.0	126.019	2134.796	0.0	71.25	92.93	121.57	273155.95	262.83	35.19	11900.489
MET	6.0	22.292	65.144	0.0	16.04	19.92	24.55	6053.41	45.48	8.61	3976.455
PHE	5.0	64.76	1154.421	16.15	46.75	54.46	63.56	216162.955	102.1	30.99	34565.75
TYR	6.0	141.91	3183.922	6.48	71.27	95.02	129.025	465375.78	329.33	34.84	15740.422
HCYS	35369.0	3.099	2.724	0.641	1.861	2.685	3.461	30.443	10.313	0.813	63.174
Pro	6.0	240.38	4139.314	0.0	137.99	166.84	200.71	400411.095	328.66	77.822	4972.429
Val	6.0	124.651	1091.825	0.0	89.27	109.48	134.26	180178.29	248.61	51.5	22573.002
BTDC	24677.0	220.602	57.781	20.313	177.377	220.542	264.098	421.514	343.35	97.189	-0.588
C0	6.0	23.756	63.869	0.0	15.55	20.44	27.14	7075.845	61.18	8.26	6198.882
C3	6.0	2.202	1.252	0.0	1.48	1.97	2.65	76.355	5.82	0.72	626.977
C4OH/C3DC	6.0	0.152	0.077	0.0	0.1	0.14	0.19	1.06	0.4	0.04	3.717
C4	6.0	0.265	0.151	0.0	0.18	0.23	0.31	10.775	0.78	0.08	673.132
C5OH/C4DC	6.0	0.186	0.068	0.0	0.14	0.18	0.22	1.82	0.39	0.07	20.522
C5	6.0	0.144	0.07	0.0	0.1	0.13	0.17	1.32	0.4	0.05	17.94
C5:1	6.0	0.009	0.007	0.0	0.01	0.01	0.01	0.2	0.03	0.0	107.226
C5DC/C6OH	6.0	0.122	0.046	0.0	0.09	0.12	0.14	1.67	0.26	0.046	47.829
C6	6.0	0.047	0.021	0.0	0.03	0.04	0.06	0.38	0.11	0.01	8.773
C6DC	6.0	0.108	0.051	0.0	0.075	0.1	0.13	3.39	0.25	0.02	496.412
C8	6.0	0.063	0.034	0.0	0.04	0.06	0.07	1.45	0.18	0.02	129.002
C8:1	6.0	0.056	0.031	0.0	0.03	0.05	0.07	0.39	0.16	0.01	5.473
C10	6.0	0.084	0.048	0.0	0.05	0.07	0.1	1.24	0.25	0.02	21.808
C10:1	6.0	0.055	0.022	0.0	0.04	0.05	0.07	0.56	0.12	0.02	15.982
C10:2	6.0	0.004	0.006	0.0	0.0	0.0	0.01	0.1	0.02	0.0	5.629
C12	6.0	0.102	0.064	0.0	0.06	0.09	0.12	2.05	0.34	0.02	37.883
C12:1	6.0	0.07	0.049	0.0	0.04	0.06	0.08	0.92	0.27	0.01	14.749
C14	6.0	0.228	0.093	0.0	0.17	0.21	0.27	4.73	0.5	0.08	197.212
C14:1	6.0	0.114	0.067	0.0	0.07	0.1	0.14	3.14	0.35	0.03	177.343
C14:2	6.0	0.023	0.011	0.0	0.02	0.02	0.03	0.49	0.05	0.01	157.072
C14-OH	6.0	0.014	0.009	0.0	0.01	0.01	0.02	0.45	0.04	0.0	191.423
C16	6.0	3.235	1.313	0.0	2.37	3.06	3.908	66.28	6.81	1.15	205.584
C16:1	6.0	0.226	0.093	0.0	0.16	0.21	0.28	1.97	0.5	0.07	6.415
C16-OH	5.0	0.02	0.015	0.0	0.01	0.02	0.02	1.92	0.05	0.01	7416.482
C16:1-OH	6.0	0.044	0.017	0.0	0.03	0.04	0.05	0.86	0.09	0.02	169.806
C18	6.0	1.052	0.443	0.0	0.78	0.99	1.25	21.955	2.19	0.42	256.211
C18:1	6.0	1.803	0.669	0.0	1.38	1.72	2.12	32.31	3.52	0.76	267.317
C18:2	6.0	0.27	0.141	0.0	0.17	0.24	0.33	2.92	0.74	0.076	10.962
C18-OH	6.0	0.013	0.01	0.0	0.01	0.01	0.02	1.14	0.03	0.0	4879.432
C18:1-OH	6.0	0.025	0.014	0.0	0.02	0.02	0.03	1.38	0.05	0.01	3060.552
C18:2OH	33231.0	0.014	0.005	0.004	0.01	0.013	0.016	0.075	0.032	0.006	14.571
C20	33231.0	0.052	0.037	0.008	0.03	0.043	0.062	0.639	0.199	0.014	38.536
C22	33231.0	0.012	0.007	0.004	0.008	0.01	0.013	0.109	0.043	0.005	35.331
C24	33231.0	0.019	0.011	0.004	0.013	0.016	0.021	0.14	0.06	0.006	23.594
C26	33231.0	0.013	0.009	0.003	0.008	0.011	0.014	0.121	0.051	0.005	33.033
SA	6.0	0.809	0.414	0.0	0.57	0.79	1.03	20.61	1.66	0.12	303.15
ADO	33231.0	0.679	0.266	0.176	0.495	0.643	0.818	3.077	1.468	0.258	6.945
D-ADO	33231.0	0.02	0.019	0.003	0.013	0.016	0.021	0.315	0.115	0.007	68.629
C20:0-LPC	33231.0	0.337	0.224	0.063	0.196	0.27	0.397	2.118	1.216	0.099	9.067
C22:0-LPC	33231.0	0.25	0.168	0.051	0.155	0.204	0.283	1.562	1.01	0.084	13.076
C24:0-LPC	33231.0	0.5	0.258	0.132	0.346	0.44	0.566	2.511	1.534	0.197	9.565
C26:0-LPC	33231.0	0.297	0.186	0.06	0.199	0.252	0.332	2.07	1.15	0.106	19.992
IRT-GSP	35485.0	18.996	7.535	8.876	13.443	17.8	22.274	47.512	47.512	8.876	3.286
TGAL	24688.0	2.32	1.788	0.0	1.059	1.887	3.118	18.931	8.331	0.0	4.869
MMA	2017.0	0.003	0.183	0.0	0.0	0.0	0.0	30.253	0.0	0.0	23172.439
EMA	35369.0	0.913	0.702	0.157	0.641	0.814	1.007	8.336	2.066	0.221	77.575
GA	35369.0	1.578	0.735	0.229	1.054	1.4	1.977	3.937	3.558	0.344	0.296
2OH GA	35369.0	16.135	7.305	5.845	10.567	14.46	21.281	39.094	33.493	6.515	0.18
3OH GA	35369.0	0.291	0.198	0.03	0.131	0.269	0.386	1.062	0.851	0.045	1.163
3OH PA	35369.0	14.557	8.865	0.305	5.07	15.593	22.643	32.387	30.403	1.866	-1.342
MCA	35369.0	10223.687	8304.851	0.0	5435.958	9482.503	13557.478	42788.038	34669.764	0.207	1.718
OROTICO	35369.0	0.397	0.72	0.0	0.234	0.314	0.415	9.147	0.884	0.013	136.867
PIVA	35369.0	0.015	0.08	0.0	0.001	0.003	0.01	0.99	0.11	0.0	137.88
2MBC	35369.0	0.115	0.068	0.0	0.066	0.097	0.145	0.356	0.333	0.015	1.696
c4-b	35369.0	0.162	0.201	0.021	0.063	0.096	0.155	1.208	0.972	0.027	9.496
c4-i	35369.0	0.151	0.093	0.007	0.095	0.134	0.179	0.76	0.407	0.032	12.947
GestationalAge	0.0	36.426	1.767	25.0	35.0	37.0	38.0	41.0	40.0	32.0	-0.12
Weight	0.0	2302.67	137.88	2000.0	2200.0	2330.0	2420.0	2499.0	2490.0	2000.0	-0.853
NTSH	0.0	2.706	4.435	0.0	1.412	2.207	3.358	415.0	9.17	0.403	5002.723

Table 5. Statistics of the observations with a weight between 2000 g and 2500 g (35527 babies)

Name	null_count	mean	std	min	25%	50%	75%	max	99%	1%	Kurtosis
ASATotal	651324.0	0.287	0.175	0.036	0.188	0.249	0.333	2.957	1.026	0.095	26.799
Ala	113.0	341.934	11038.257	0.0	205.73	248.34	302.23	8401024.23	527.97	126.72	480822.092
Arg	113.0	9.885	21.311	0.0	4.77	8.077	12.84	5756.92	33.55	1.0	27669.777
Cit	113.0	14.299	28.637	0.0	10.48	13.26	16.6	9043.766	30.49	4.38	25763.555
Glu	651324.0	249.736	1167.583	23.454	198.576	235.234	280.22	252238.877	441.568	130.896	46430.007
Gly	113.0	545.912	7610.805	0.0	339.09	413.81	504.1	943330.74	898.93	166.71	6330.751
Leu/Ile/Pro-OH	113.0	182.503	2739.295	0.0	118.89	139.76	164.7	409473.065	253.5	79.47	7492.865
Orn	113.0	125.166	1853.378	0.0	77.01	97.33	125.01	504365.6	268.17	43.55	25142.402
MET	113.0	19.679	35.664	0.0	15.28	18.69	22.49	7773.35	35.15	8.74	10847.831
PHE	111.0	59.661	627.636	0.0	47.64	54.69	62.95	362390.198	93.1	33.77	183574.406
TYR	111.0	110.837	1764.938	0.0	69.3	86.511	109.34	354293.85	221.06	39.99	18844.445
HCYS	695488.0	3.231	2.474	0.0	2.177	2.82	3.764	76.302	9.706	0.844	337.22
Pro	113.0	224.157	3036.979	0.0	149.98	174.78	204.66	527548.27	318.72	102.129	6732.61
Val	113.0	509.427	285179.617	0.0	108.36	129.64	155.3	238238232.0	249.22	68.85	697831.746
BTd	479344.0	245.461	56.596	3.403	205.955	249.976	288.215	460.641	354.753	111.472	-0.378
C0	113.0	19.022	38.822	0.0	12.89	16.89	22.24	8278.81	45.68	6.81	9077.626
C3	113.0	2.105	2.021	0.0	1.46	1.92	2.52	1010.575	5.24	0.73	177829.286
C4OH/C3DC	114.0	0.2	0.089	0.0	0.14	0.19	0.25	3.63	0.47	0.06	8.905
C4	113.0	0.247	0.137	0.0	0.17	0.22	0.29	32.68	0.7	0.07	4946.763
C5OH/C4DC	114.0	0.207	0.092	0.0	0.16	0.2	0.241	26.76	0.42	0.08	12894.085
C5	113.0	0.111	0.064	0.0	0.08	0.1	0.13	10.92	0.3	0.04	7362.983
C5:1	113.0	0.009	0.007	0.0	0.006	0.01	0.01	1.19	0.03	0.0	2371.094
C5DC/C6OH	114.0	0.131	0.05	0.0	0.1	0.12	0.15	15.22	0.27	0.05	11680.347
C6	113.0	0.047	0.023	0.0	0.03	0.04	0.06	4.93	0.11	0.01	5319.965
C6DC	113.0	0.127	0.056	0.0	0.09	0.12	0.16	15.53	0.28	0.03	8365.201
C8	113.0	0.064	0.1	0.0	0.04	0.06	0.074	37.58	0.18	0.02	81092.019
C8:1	113.0	0.033	0.021	0.0	0.02	0.03	0.04	1.26	0.11	0.0	56.499
C10	113.0	0.096	0.052	0.0	0.06	0.084	0.114	11.0	0.27	0.03	2910.925
C10:1	113.0	0.048	0.019	0.0	0.04	0.05	0.06	1.22	0.11	0.02	103.127
C10:2	113.0	0.003	0.005	0.0	0.0	0.0	0.006	0.87	0.02	0.0	1909.363
C12	113.0	0.126	0.069	0.0	0.08	0.11	0.15	11.175	0.36	0.04	968.35
C12:1	113.0	0.093	0.056	0.0	0.06	0.08	0.11	2.01	0.291	0.02	14.235
C14	113.0	0.241	0.087	0.0	0.18	0.23	0.28	10.805	0.51	0.1	412.933
C14:1	113.0	0.133	0.081	0.0	0.09	0.12	0.16	30.086	0.37	0.04	27697.317
C14:2	113.0	0.02	0.012	0.0	0.01	0.02	0.02	5.07	0.05	0.01	46502.796
C14-OH	113.0	0.018	0.01	0.0	0.01	0.02	0.02	0.63	0.05	0.0	44.661
C16	113.0	3.716	1.379	0.0	2.84	3.54	4.38	108.655	7.36	1.59	384.527
C16:1	113.0	0.257	0.092	0.0	0.19	0.246	0.31	11.026	0.521	0.1	275.162
C16-OH	113.0	0.024	0.011	0.0	0.02	0.02	0.03	2.34	0.06	0.01	3265.006
C16:1-OH	113.0	0.045	0.015	0.0	0.03	0.04	0.05	0.62	0.09	0.02	13.988
C18	113.0	1.091	0.417	0.0	0.83	1.03	1.28	56.17	2.21	0.48	908.261
C18:1	113.0	1.721	0.625	0.0	1.34	1.64	2.01	44.175	3.3	0.79	568.54
C18:2	114.0	0.162	0.084	0.0	0.11	0.14	0.2	10.685	0.44	0.05	600.728
C18-OH	113.0	0.016	0.007	0.0	0.01	0.017	0.02	1.11	0.033	0.0	1498.947
C18:1-OH	114.0	0.026	0.01	0.0	0.02	0.02	0.03	1.95	0.05	0.01	3309.221
C18:2OH	651324.0	0.011	0.004	0.001	0.008	0.01	0.012	0.246	0.024	0.005	267.599
C20	651324.0	0.05	0.03	0.003	0.032	0.042	0.058	0.497	0.166	0.017	24.32
C22	651324.0	0.011	0.006	0.002	0.008	0.01	0.012	0.106	0.039	0.005	29.433
C24	651324.0	0.019	0.009	0.001	0.013	0.017	0.021	0.265	0.058	0.008	30.3
C26	651324.0	0.012	0.008	0.001	0.008	0.01	0.014	0.235	0.046	0.005	53.594
SA	112.0	0.805	0.391	0.0	0.57	0.78	1.02	30.62	1.66	0.12	386.402
ADO	651324.0	0.588	0.223	0.062	0.434	0.56	0.707	10.889	1.231	0.232	127.68
D-ADO	651324.0	0.019	0.017	0.0	0.013	0.016	0.021	0.438	0.103	0.007	85.385
C20:0-LPC	651324.0	0.286	0.192	0.024	0.169	0.23	0.335	2.947	1.061	0.091	13.144
C22:0-LPC	651324.0	0.241	0.155	0.017	0.154	0.199	0.27	3.602	0.914	0.087	20.323
C24:0-LPC	651324.0	0.518	0.264	0.04	0.362	0.454	0.584	10.897	1.539	0.202	60.617
C26:0-LPC	651324.0	0.296	0.172	0.028	0.204	0.256	0.331	2.649	1.066	0.113	22.877
IRT-GSP	697169.0	19.743	10.68	0.0	12.514	17.314	23.666	115.88	57.584	5.24	9.897
TGAL	479543.0	2.005	2.015	0.0	0.844	1.557	2.652	470.459	7.978	0.0	13600.575
MMA	40991.0	0.004	0.85	0.0	0.0	0.0	0.0	577.73	0.0	0.0	352611.265
EMA	695488.0	1.212	1.602	0.0	0.653	0.897	1.261	35.006	6.807	0.225	159.333
GA	695488.0	1.73	1.527	0.0	1.175	1.578	2.078	65.194	4.563	0.368	1164.591
2OH GA	695488.0	17.859	8.845	0.0	10.802	16.345	23.371	67.636	43.325	4.888	1.218
3OH GA	695488.0	0.369	0.338	-0.789	0.158	0.311	0.528	10.843	1.094	0.029	365.193
3OH PA	695488.0	14.612	8.964	0.202	5.313	15.28	21.965	60.09	34.931	1.576	-0.237
MCA	695488.0	9807.239	9817.433	0.0	4743.558	9010.022	13499.174	223457.997	31605.118	0.132	174.093
OROTICO	695488.0	0.448	1.334	0.0	0.246	0.337	0.468	39.732	1.046	0.042	444.786
PIVA	695488.0	0.006	0.031	0.0	0.001	0.002	0.004	0.969	0.12	0.0	427.884
2MBC	695488.0	0.08	0.043	0.0	0.053	0.072	0.096	0.606	0.236	0.012	19.586
c4-b	695488.0	0.184	0.239	0.0	0.062	0.097	0.176	2.433	1.127	0.014	13.343
c4-i	695488.0	0.216	3.328	0.0	0.081	0.117	0.17	167.685	0.65	0.017	2500.991
GestationalAge	0.0	39.111	1.288	23.0	38.0	39.0	40.0	43.0	41.0	36.0	0.72
Weight	0.0	3318.163	403.123	2500.0	3030.0	3300.0	3580.0	5000.0	4340.0	2540.0	-0.071
NTSH	7.0	2.487	3.824	0.0	1.356	2.056	3.072	566.0	8.12	0.416	5512.588

Table 6. Statistics of the observations with a weight higher than 2500 g (698052 babies)

Weight	Size	Positives	Percentage
< 1500	5966	5	0.08
1500 - 2000	9172	0	0.00
2000 - 2500	35527	14	0.04
> 2500	698051	179	0.03

Table 7. Table showing the number of positive cases among infants in each weight range.

References

1. Bickel, H. *et al.* Neonatal mass screening for metabolic disorders: summary of recent sessions of the committee of experts to study inborn metabolic diseases, public health committee, council of europe. *Eur. J. Pediatr.* **137**, 133–139, <https://doi.org/10.1007/BF00441305> (1981).
2. Mütze, U. *et al.* Long-term outcomes of individuals with metabolic diseases identified through newborn screening. *Pediatr.* **146**, <https://doi.org/10.1542/peds.2020-0444> (2020).
3. Oltman, S. P. *et al.* Newborn metabolic vulnerability profile identifies preterm infants at risk for mortality and morbidity. *Pediatr. research* **89**, 1405–1413, <https://doi.org/10.1038/s41390-020-01148-0> (2021).
4. Cassio, A. *et al.* The italian screening program for primary congenital hypothyroidism: actions to improve screening, diagnosis, follow-up, and surveillance. *J. Endocrinol. Investig.* **36**, 195–203, <https://doi.org/10.3275/8849> (2013).
5. Beyan, H. *et al.* Guthrie card methylomics identifies temporally stable epialleles that are present at birth in humans. *Genome research* **22**, 2138–2145, <https://doi.org/10.1101/gr.134304.111>. (2012).
6. Fingerhut, R. & Torresani, T. Evaluation of the genetic screening processor (gsp™) for newborn screening. *Anal. Methods* **5**, 4769–4776, <https://doi.org/10.1039/C3AY40593A> (2013).
7. Glish, G. L. & Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nat. reviews drug discovery* **2**, 140–150, <https://doi.org/10.1038/nrd1011> (2003).
8. la Marca, G. *et al.* Current state and innovations in newborn screening: Continuing to do good and avoid harm. *Int. J. Neonatal Screen.* <https://doi.org/10.3390/ijns9010015> (2023).
9. Ospedale dei Bambini Vittore Buzzi. Extended Lombardy's Neonatal Screening Dataset. **Zenodo**, [10.5281/zenodo.164111492025](https://doi.org/10.5281/zenodo.164111492025).
10. Ridzuan, F. & Zainon, W. M. N. W. A review on data cleansing methods for big data. *Procedia Comput. Sci.* **161**, 731–738, <https://doi.org/10.1016/j.procs.2019.11.177> (2019).
11. Kaabachi, B. *et al.* Can we trust synthetic data in medicine? a scoping review of privacy and utility metrics. *medRxiv* 2023–11, <https://doi.org/10.1101/2023.11.28.23299124> (2023).
12. Lu, Y. *et al.* Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062* <https://doi.org/10.48550/arXiv.2302.04062> (2023).
13. Berardo C. & Vasco A. & Mauri A. & Lucchi S. & Cappelletti L. & Saielli L. & Rizzetto M. & Biganzoli D. & Montrasio C. & Postorivo D. & et al. Expanded Newborn Screening in Italy: The First Report of Lombardy Region. *International Journal of Neonatal Screening*. 2025; 11(2):31. <https://doi.org/10.3390/ijns11020031>
14. Vesin, A. *et al.* Reporting and handling missing values in clinical studies in intensive care units. *Intensive care medicine* **39**, 1396–1404, <https://doi.org/10.1007/s00134-013-2949-1> (2013).
15. Janssen, K. J. *et al.* Missing covariate data in medical research: to impute is better than to ignore. *J. clinical epidemiology* **63**, 721–727, <https://doi.org/10.1016/j.jclinepi.2009.12.008> (2010).
16. Guan, N. C. & Yusoff, M. S. B. Missing values in data analysis: ignore or impute? *Educ. Medicine J.* **3**, <https://doi.org/10.5959/eimj.3.1.2011.or1> (2011).
17. Cismondi, F. *et al.* Missing data in medical databases: Impute, delete or classify? *Artif. intelligence medicine* **58**, 63–72, <https://doi.org/10.1016/j.artmed.2013.01.003> (2013).
18. Zeng, D., Xie, D., Liu, R. & Li, X. Missing value imputation methods for tcm medical data and its effect in the classifier accuracy. In *2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom)*, 1–4, <https://doi.org/10.1109/HealthCom.2017.8210844> (IEEE, 2017).
19. Cheng, C.-H., Chang, J.-R. & Huang, H.-H. A novel weighted distance threshold method for handling medical missing values. *Comput. Biol. Medicine* **122**, 103824, <https://doi.org/10.1016/j.combiomed.2020.103824> (2020).
20. Afkanpour, M., Hosseinzadeh, E. & Tabesh, H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Med. Res. Methodol.* **24**, 188, <https://doi.org/10.1186/s12874-024-02310-6> (2024).
21. Ieee recommended practice for the quality management of datasets for medical artificial intelligence. *IEEE Std 2801-2022* 1–31, <https://doi.org/10.1109/IEEESTD.2022.9812564> (2022).