OBUS – Data Description

Overview

The Global Health Labs (GHL) machine learning team used datasets from the Fetal Age Machine Learning Initiative (FAMLI) to train and evaluate a starter package of AI models for AI-enabled obstetric ultrasound (OBUS). The goal of the OBUS-GHL project is to develop AI models that can assist in the interpretation of obstetric ultrasound videos. The project focuses on four models, each for a key obstetric feature: GA (gestational age estimation), FP (fetal presentation classification), EFW (fetal weight estimation), and TWIN (multiple gestation classification).

FAMLI is spearheaded by the University of North Carolina at Chapel Hill (UNC), with the goal of expanding ultrasound access in low-income communities. It is funded by the Gates Foundation, with complementary resources from UNC, the National Institutes of Health, and the Butterfly Network, Inc. The FAMLI datasets consist of obstetric ultrasound videos collected by UNC and partners at clinics and hospitals in North Carolina and Zambia. They comprise thousands of patients and exams, and hundreds of thousands of ultrasound videos. Prospective data was collected in phases, beginning in September 2018 at two sites in Chapel Hill, North Carolina and in January 2019 at four sites in Lusaka, Zambia.

In the initial phase of data collection, UNC enrolled women over the age of 18, with a singleton intrauterine uncomplicated pregnancy. Obstetric ultrasound videos of the gravid abdomen were collected at the initial exam and subsequent follow-up exams. The initial visit was used to collect demographic information and to determine the best estimate of gestational age. At each visit, obstetric features of interest were observed/measured and recorded. In later phases of data collection, the inclusion criteria were expanded to include multiple gestation pregnancies as well as pregnancies with various risk factors to allow for development of obstetric feature algorithms targeting those conditions.

Generally, there are at least six blind video sweeps per exam: three vertical (cranio-caudal) sweeps and three horizontal (lateral) sweeps. In practice, there are between two and fifty videos per exam (including non-blind sweeps) in the FAMLI dataset. Each sweep should last about 10 seconds. In practice, the video lengths are between 2 sec and 70 secs. The vertical sweeps start at the pubis and end at the level of the uterine fundus, with the probe indicator facing the maternal right. The middle one goes through the belly button and is labelled M; the meridians to the maternal right are labelled, (R0), R1, R2, etc.; those to the left, (L0), L1, L2, etc. The horizontal sweeps start just above the pubis, sweeping from the maternal left to the maternal right uterine borders, and progress in cephalad direction until the uterine fundus is reached. The probe indicator faces superiorly. The lowest sweep is labelled C1, the next superior one C2, etc. Figure 1 shows a diagram of these vertical and horizontal blind sweeps. A video of the blind sweep procedure may be found here.

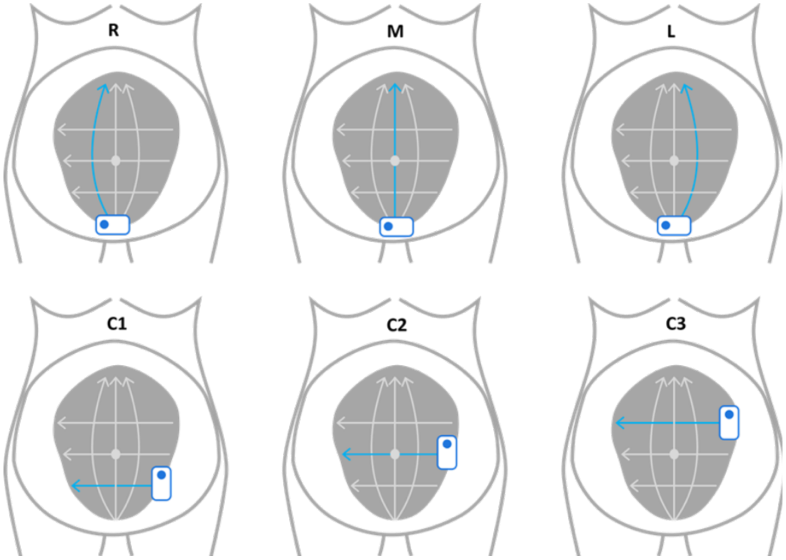

Figure 1. Blind sweep definitions

For model development, GHL used the four main sets of FAMLI data described below.

- **FAMLI2_enrolled**: Participants enrolled 2018-09 through 2022-05.
- **NEJM**: A highly curated subset of FAMLI2_enrolled data used by UNC to develop their own AI models, which they published in the New England Journal of Medicine Evidence [1]
- **FAMLI2**: Participants enrolled 2022-06 through 2023-05.
- **FAMLI3**: Participants enrolled beginning 2023-06. The OBUS-GHL code base [2] was developed and tested with data collected through 2025-04, but should work with subsequent data collected, if the protocols and tables remain consistent.

NEJM Dataset

GHL trained and evaluated models for the gestational age (GA) and fetal presentation (FP) models on the NEJM dataset, using the training, validation, and testing splits defined by UNC. The counts of videos and exams in the NEJM dataset are shown in Table 1, which lists statistics for the five splits of the NEJM data separately, as well as for the entire dataset in the final three rows. The table shows statistics for three kinds of exams: those acquired by an expert with a high-cost device (GE or Sonosite); those acquired by an expert with a low-cost device (Butterfly); and those acquired by a novice with a low-cost device. The novice exams were only collected in the Zambian clinics.

	User, Manufacturer	Expert, Other (GE or Sonosite)	Expert, Butterfly	Novice, Butterfly	All users, devices
Training set	# patients	2,783	2,095	192	5,070
	# exams	4,687	3,185	195	8,067
	# videos	50,005	33,876	1,202	85,083
Tuning set	# patients	697	527	46	1,270
	# exams	1,161	784	47	1,992
	# videos	12,249	8,220	315	20,784
Main testing set	# patients	716	0	0	716
	# exams	1,278	0	0	1,278
	# videos	12,053	0	0	12,053
IVF testing set	# patients	47	0	0	47
	# exams	79	0	0	79
	# videos	518	0	0	518
Novice testing set	# patients	0	0	129	129
	# exams	0	0	147	147
	# videos	0	0	992	992
All NEJM	# patients	4,243	2,622	367	7,232
	# exams	7,205	3,969	389	11,563
	# videos	74,825	42,096	2,509	119,430

Table 1. Ultrasound patients, exams, and videos by user, device in the NEJM dataset

The STARD study flow chart describing data allotments for Machine Learning purposes, including number of exams in the NEJM subset of FAMLI2_enrolled, is shown in Figure 2.

Combined Dataset

The combined dataset comprises the FAMLI2_enrolled, FAMLI2, and FAMLI3 datasets. To combine the FAMLI2_enrolled, FAMLI2, and FAMLI3 datasets, data harmonization had to be applied to reconcile the differences in metadata between the constituent sets. The data harmonization scheme is described in detail in the README document [2].

As model development occurred during data collection, several different versions of the dataset were used as models were trained and evaluated. Table 2 shows these versions of the combined dataset along with a brief description. The v7, v8, and v9.0 datasets were interim stopping points on the pathway towards the final datasets, v9.2 and v9.3.

GHL trained and evaluated the fetal weight (EFW) and multiple gestation (TWIN) models on the combined dataset. The EFW and TWIN models, delivered as the starter package, used only the latest combined dataset, v9.3, whose details are described herein. For the combined dataset, GHL defined the training, validation, and testing splits separately for each obstetric feature and independently of UNC.

Name	Description
v7	Ingestion run in 2024-09 with FAMLI3 data up to ~2024-08
v8	v7 bugs fixed, data added to v7 with FAMLI3 data up to ~2024-10
v9.0	Fresh ingestion with FAMLI3 data up to ~2025-01
v9.2	Additional data on top of v9.0 with FAMLI3 data up to 2025-03
v9.3	Fresh ingestion with FAMLI3 data up to 2025-04

Table 2. Summary of combined dataset evolution.

The counts of videos and exams in the NEJM dataset are shown in Table 3, which lists statistics for the three subsets of the combined data separately, as well as for the entire dataset in the final three rows. The table shows statistics for five manufacturers of ultrasound devices, with GE and Butterfly comprising a number of different model devices (counts are not broken down by model).

	Manufacturer	GE	Sonosite	Butterfly	Clarius	EchoNous	All devices
FAMLI2_ enrolled	# patients	4,536	2,430	5,171	651	21	12,809
	# exams	7,783	3,646	9,122	1,268	22	21,841
	# videos	176,610	73,651	137,642	18,056	205	406,164
FAMLI2	# patients	321	0	315	0	19	655
	# exams	537	0	1,261	0	20	1,818
	# videos	13,958	0	12,264	0	196	26,418
FAMLI3	# patients	2,105	0	2,208	0	0	4,313
	# exams	3,993	0	6,898	0	0	10,891
	# videos	112,779	0	109,087	0	0	221,866
Combined	# patients	6,962	2,430	7,694	651	40	17,777
	# exams	12,313	3,646	17,281	1,268	42	34,550
	# videos	303,347	73,651	258,993	18,056	401	654,448

Table 3. Ultrasound patients, exams, and videos by device in the combined dataset

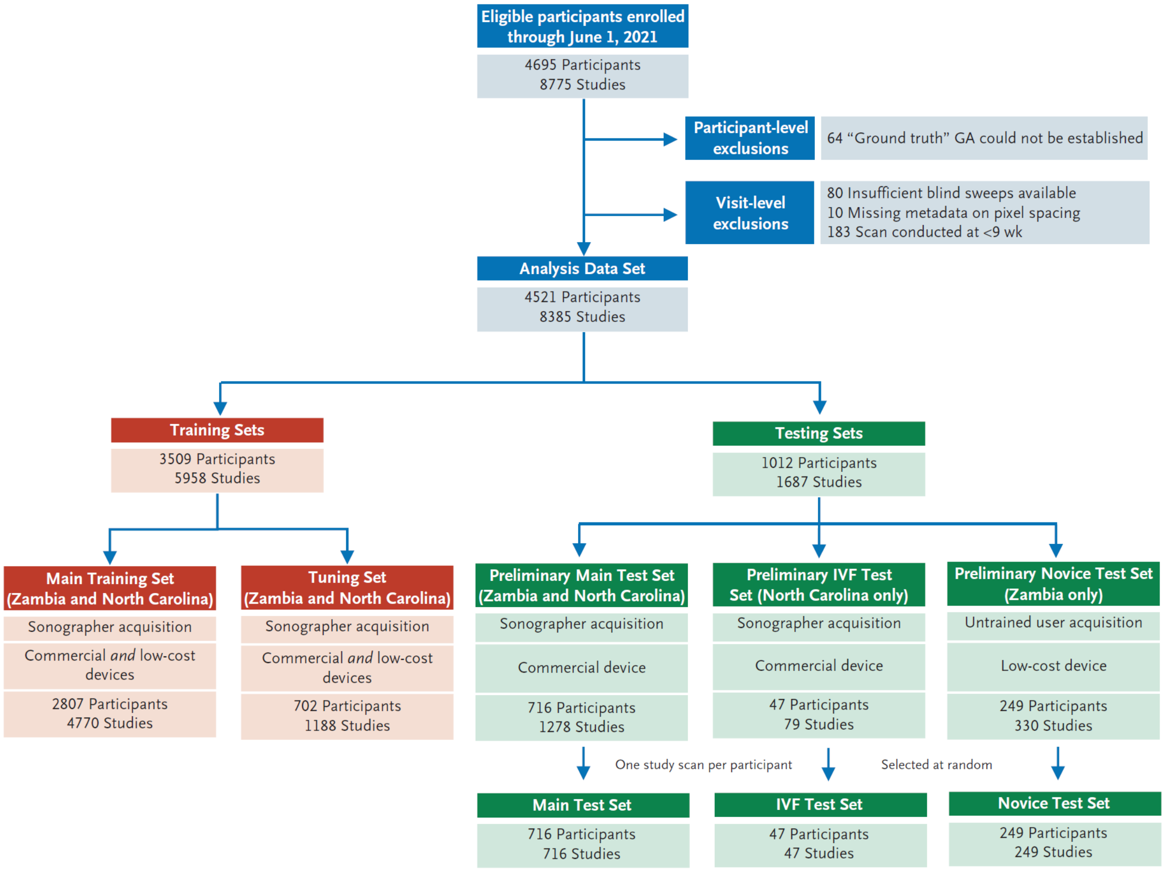

Figure 2. NEJM Data Collection Study Flow Chart (STARD) [1]

References

[1] T. Pokaprakarn, J. S. Stringer and et al., "AI estimation of gestational age from blind ultrasound sweeps in low-resource settings," *NEJM Evidence,* vol. 1, no. 5, p. EVIDoa2100058, 2022.

[2] "GHL OBUS GitHub Repository," [Online]. Available: https://github.com/Global-Health-Labs/OBUS-GHL.