

# OBUS – Multiple Gestation Data

## Overview

The extended FAMI dataset (FAMI2\_enrolled, FAMI2, and FAMI3) was used to train and evaluate multiple gestation (MG) models. Different versions of the combined dataset are described in [\[0.1 OBUS Data Description\]](#). The division of data into distributions also evolved during development. In particular, due to the very limited number of multiple gestation exams, the final model development split the data into an evaluation set and a development set, with the development set further split into five subsets to enable 5-fold cross validation. Earlier exploration, on the other hand, was based on a single train/val split within the development set. The results shown in [\[4.3 Multiple Gestation Evaluation Report\]](#) were based on models developed using the 5-fold cross-validation development set of the v9.3 dataset (see Table 2 in [\[0.1 OBUS Data Description\]](#)). We will thus only describe this final dataset and its construction here.

## Label Definition

The first step in creating the distribution is to identify all the multiple gestation patients and to select singleton patients that have both sweep tag and gestational age information available. MG ground truth labels were occasionally inconsistent between different metadata sources; thus, third-party sonographers and the UNC development team had to be consulted to resolve these inconsistencies. There were 105 twin patients and one triplet patient in the combined dataset, for a total of 106 MG patients. The MG target was therefore chosen to be any pregnancy with 2 or more babies, most of which are twins. Singleton pregnancies were assigned the label 0, while multiple gestations ( $\geq 2$ ) were assigned the label 1. The MG labels for these patients were encoded in an auxiliary ground truth file, which is explained in detail in the repository README [1].

## Data Distributions

The next step was to create data splits at the patient level. Due to the very limited number of multiple gestation exams, the data were split into a holdout (testing) set and a development set. The holdout data is not used in model training or hyperparameter optimization or model selection. Its sole purpose is to measure final model performance.

The development set was then further split into five parts (at the patient level) for 5-fold cross-validation. Singleton and MG patients were handled differently due to the extreme rarity of MG data.

The MG patients were split manually into testing, Fold0, Fold1, Fold2, Fold3, and Fold4, while trying to keep the exam-level distribution across gestational age the same for all 6 subsets. A master spreadsheet was used as a dashboard for this manual operation.

Singleton patients were subsampled randomly from among all three datasets FAML12\_enrolled, FAML12, and FAML13. There was a two-fold purpose for this: (1) to limit the number of samples so that singleton data doesn't overwhelm MG data; and (2) so that the exam-level gestational age distribution matched that of the entirety of the datasets. In contrast, all of the multiple gestation data present in the combined datasets were used for model development and evaluation. For the 5-fold cross-validation splits, a balancing algorithm was used to flatten the exam-level gestational age distribution via over-sampling. (A limit was set on the amount of oversampling allowed in each gestational age bin.)

One additional split, a 'calibration' set, was made up of singleton exams that weren't used in other splits. This allows comparing performance of the different models resulting from the cross-validation process on the same dataset, albeit limited to only singletons, prior to evaluating on the test set. It can be used, along with the validation sets in each fold, to select the 'best' model and model thresholds, keeping the test set as a true holdout. The statistics and uses of the 5-fold splits, the test set, and the calibration set are shown in Table 1.

<b>Fold</b>	<b>Patients</b>	<b>Exams</b>	<b>Videos</b>	<b>Purpose</b>
Test - Single - Multiple	414 392 22	1,371 1,333 38	16,623 16,040 583	Estimating final performance
0 - Single - Multiple	218 201 17	768 707 61	13,254 11,968 1,286	Validation for CV Fold0 Training for CV Folds 1, 2, 3, 4
1 - Single - Multiple	217 201 16	881 820 61	13,990 12,602 1,388	Validation for CV Fold1 Training for CV Folds 0, 2, 3, 4
2 - Single - Multiple	219 201 18	774 713 61	13,238 11,854 1,384	Validation for CV Fold2 Training for CV Folds 0, 1, 3, 4

3	218	788	12,728	Validation for CV Fold3
- Single	201	732	11,536	Training for CV Folds 0, 1, 2, 4
- Multiple	17	56	1,192	
4	217	749	12,465	Validation for CV Fold4
- Single	201	689	11,221	Training for CV Folds 0, 1, 2, 3
- Multiple	16	60	1,244	
Calibration (Single only)	510	589	9,710	Comparing cross validation models

*Table 1. Summary of data splits / folds*