From Benchmarks to Blind Spots: Intersectionality Gap in Fetal Ultrasound Data

Tommaso Ruga[1,2][0009−0006−3662−6955], Genoveva Vargas-Solar[3][0000−0001−9545−1821], and Ester Zumpano[1,2][0000−0003−1129−3737]

[1] DIMES - University of Calabria, Rende(CS), Italy
{tommaso.ruga,e.zumpano}@dimes.unical.it
[2] CNR-NANOTEC National Research Council, Rende(CS), Italy
[3] CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, Villeurbane, France
genoveva.vargas-solar@cnrs.fr

Abstract. This paper provides a comparative, situated review of representative fetal ultrasound datasets across regions and pregnancy contexts, and we contrast their operationalization of "pregnancy status" with non-imaging monitoring practices that structure real-world antenatal pathways (e.g., symphysis, fundal height trajectories, perceived fetal movements, and palpation-based assessment). We show that key information required for equity-aware research is systematically missing or under-specified, including longitudinal linkage across gestational time, care-pathway signals and referral reasons, and socio-economic and cultural context needed to interpret missingness and dataset shift. We argue that these omissions constrain the ability to audit whether measurement error, follow-up intensity, and model performance are socially patterned, and they risk producing benchmarks that overestimate generalisability while masking inequities.

Keywords: Fetal ultrasound datasets · intersectionality · pregnancy follow-up · dataset shift · responsible AI · data governance

1 Introduction

Prenatal care is now widely framed as "data-driven." Clinics rely on medical records, registries, digital follow-up tools, and, increasingly, AI. Ultrasound sits at the core of this ecosystem. It supports routine measures such as head circumference, abdominal circumference, and femur length. Global initiatives like INTERGROWTH-21st promote shared growth standards. However, shared standards do not mean shared data. Pregnancy data are produced under very different conditions across countries, hospitals, and social groups. These conditions are shaped by income, migration, geography, discrimination, and access to care.

Artificial Intelligence (AI) can amplify these differences. Models learn patterns from the data environment, not only from fetal biology. They pick up device settings, operator practices, local protocols, and documentation habits. When these change, performance can drop—often in settings already affected by

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

limited resources, short appointments, or reduced follow-up. As a result, a model may look strong on benchmarks but fail in real deployments and widen gaps in care. This risk is visible in predictive work on outcomes such as preterm birth or hypertensive disorders, where reviews often report weak validation and poor transportability. Bias studies also show that common proxies, like healthcare use, can reflect unequal access rather than true need, which may underestimate risk for people facing barriers to care [29].

Yet studying these issues with an intersectional lens remains difficult. Many datasets do not include the social variables needed to examine differences across groups. When such variables exist, they are often not comparable across countries or are restricted by governance rules. Monitoring infrastructures also vary widely, from European initiatives to DHIS2-based systems and North American surveillance. In many cases, "good data" is defined by reporting indicators, while lived experience and cultural context are reduced to crude proxies—or not captured at all.

This paper addresses this gap by focusing on fetal ultrasound datasets as a particularly influential substrate for contemporary obstetric AI. We argue that the field lacks integrated studies that jointly examine (i) the quantitative properties of publicly available fetal ultrasound datasets (coverage, composition, and dataset shift drivers), and (ii) the socio-economic and cultural conditions of pregnancy follow-up that shape which pregnancies are measured, how measurements are produced, and how missingness is socially patterned. We therefore conduct a comparative, intersectionality-aware analysis of publicly available fetal ultrasound datasets and contrast them with non-imaging monitoring practices that structure real-world pathways to ultrasound. By doing so, we aim to clarify what is missing for an intersectional view of fetal evolution during pregnancy, and to motivate dataset documentation and design requirements that better support responsible, equity-aware research and deployment.

The reminder of the paper is organized as follows. Section 2 introduces related work on pregnancy studies datasets and works addressing the problem efficiently and objectively relying on data analysis. Section 3 proposes a comparative landscape of fetal ultrasound datasets and non-imaging monitoring practices. Section 4 describes the study we propose to map the gap and determine who is missing from datasets representation. Finally, section 5 concludes the paper and discusses future work.

2 Dataset Bias and the Generalization Problem: Current Evidence

Pregnancy follow-up is increasingly "data-driven" through the combination of clinical records, population registries, routine health information systems, and commercial digital tools. Yet, across these infrastructures, what is collected (and what is not) is shaped by institutional priorities, legal regimes, and implicit assumptions about whose pregnancy experiences are legible in data.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

Perinatal surveillance infrastructures and registries. In Europe, cross-national monitoring of maternal and newborn outcomes relies on heterogeneous national sources (civil registration, medical birth registers, hospital discharge data, congenital anomaly registers, and audit systems), which complicates comparability and limits the granularity of inequality analyses [11,39]. The Euro-Peristat network has documented persistent methodological challenges (e.g., differing registration criteria and definitions, partial coverage, and missingness for rarer events) and has recently advanced toward more scalable solutions, including open protocols and common data models for federated analysis of national birth data [40,28]. These efforts improve the feasibility of international comparisons, but they still largely inherit the variable availability and sociopolitical constraints of national systems, particularly for sensitive or contested dimensions of social stratification.

Northern European medical birth registers are described as uniquely valuable for longitudinal research due to stable identifiers and linkability across health and administrative data, enabling follow-up beyond delivery [20]. However, even in these "gold standard" contexts, intersectional analyses can be constrained by how migration, ethnicity, disability, or social position are (or are not) encoded and by differential data access governance. In the United Kingdom, confidential surveys and audit infrastructures such as MBRRACE-UK [4] provide deep interpretive insight into maternal deaths and severe morbidity, complementing routine data with structured review and preventability analyses While analytically powerful, such systems often prioritize clinical narratives and service pathways; they do not automatically resolve cross-country comparability issues, and their categories may not align with those used elsewhere.

In North America, pregnancy-related surveillance includes population-based surveys and linked vital records. The U.S. Pregnancy Risk Assessment Monitoring System (PRAMS) is frequently cited as a model for capturing behaviors and experiences before, during, and shortly after pregnancy via mixed-mode sampling anchored in birth certificates, with explicit capacity to study disparities through stratified sampling [34]. Yet, PRAMS is not designed as a universal individual-level follow-up infrastructure and remains conditioned by jurisdictional participation, variable item availability, and the politics of measuring race/ethnicity, insurance, and social determinants. Together, these registry and surveillance ecosystems show that "better data" initiatives tend to emphasize harmonization of indicators and reporting pipelines, while leaving unresolved the deeper question of which lived experiences are structurally absent from the variables themselves.

Digital follow-up systems and routine health information platforms. In many settings in Latin America, pregnancy follow-up and perinatal reporting have been supported by long-standing clinical registry and quality-improvement infrastructures such as PAHO/CLAP's Perinatal Information System (SIP), de-

[4] *MBRRACE-UK Maternal Report 2023*. Available at: `https://www.npeu.ox.ac.uk/assets/downloads/mbrrace-uk/reports/maternal-report-2023/MBRRACE-UK_Maternal_Compiled_Report_2023.pdf` [Accessed 20 Feb 2026].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

signed to consolidate pregnancy and birth records and to support local reporting and program monitoring [35,26]. SIP demonstrates how a region-wide tool can stabilize data practices across diverse health systems, but it also illustrates a common tension: standardization for comparability may under-specify intersectional context (e.g., locally meaningful social categories, informal care trajectories, or structural violence exposures) that strongly influence outcomes.

Across parts of Africa and Asia, pregnancy follow-up is frequently mediated through routine health information systems (RHIS) and digital platforms such as DHIS2. DHIS2's wide adoption has enabled more timely reporting and the possibility of data-driven management, including for maternal and neonatal indicators [2,41]. However, empirical studies repeatedly highlight gaps in completeness, consistency, and the alignment of facility registers with digital aggregates, underscoring that data quality is produced through organizational practices (supervision, feedback loops, workload, infrastructure reliability) rather than software alone [2,23]. These challenges are consequential for equity: if missingness or reporting error is socially patterned (by geography, facility type, language, or stigma), then "data-driven" follow-up can systematically under-serve groups already facing barriers to care.

Large-scale beneficiary tracking systems also exist, such as India's Mother and Child Tracking System (MCTS), which aimed to support service delivery planning by tracking maternal and child health beneficiaries. Assessments of MCTS document both its promise and persistent implementation challenges (data completeness, timeliness, usability, and integration with frontline workflows), again indicating that follow-up capacity depends on socio-technical fit rather than digitization alone [14]. Overall, these infrastructures show a recurrent pattern: pregnancy follow-up datasets are assembled across multiple layers (clinical records, registers, reporting aggregates, and program databases), and the "gaps" often emerge at the seams, where categories do not match, identifiers do not link, and local meaning is lost in aggregation.

Predictive modelling in obstetrics and data dependencies. A rapidly growing body of work applies statistical learning and machine learning to predict pregnancy complications (e.g., preterm birth, hypertensive disorders, hemorrhage) using EHRs, claims, registries, and sometimes wearable/app data. Systematic reviews emphasize uneven model quality, limited external validation, and recurring risk-of-bias concerns (including outcome definition drift, missingness handling, and dataset shift across settings) [15,21]. Even when performance appears strong within a single institution, portability across countries and care pathways is often weak because models entangle local clinical practice patterns, coding regimes, and access-to-care structures.

Work on algorithmic bias in healthcare provides an important warning for pregnancy contexts: predictive systems can reproduce structural inequities when proxies (e.g., cost, utilization, or recorded diagnoses) stand in for need in ways that reflect unequal access to care [25]. For pregnancy follow-up, this implies that models trained on datasets where marginalized groups experience delayed diagnosis, under-documentation, or differential treatment may systematically

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

underestimate risk or misallocate resources. Yet, much of the obstetric ML literature still treats "data" as a neutral substrate, rather than as a situated trace of unequal systems. This motivates closer attention to how pregnancy datasets are produced, categorized, and governed across settings, before claiming generalizable "data-driven" follow-up.

Intersectional, situated views on pregnancy data gaps. Feminist STS and critical data studies treat dataset gaps as outcomes of power and governance, not just technical noise. Situated knowledge shows that data are never viewpoint-free [17]. Intersectionality shows that harm and exclusion emerge from combined structures such as gender, race, class, migration, and disability [7,8]. Classification systems and standards shape what becomes visible and actionable, and what is left out [3,16]. Data feminism turns these ideas into practice by asking researchers to examine power, value embodied knowledge, and read missingness as a political signal [10]. Applied to pregnancy follow-up, this means that missing intersectional variables, or reducing them to crude proxies, is rarely accidental. It often reflects legal constraints, administrative routines, past harms of categorization, and the priorities of health systems and platforms. Cross-country work adds another challenge: variables that seem equivalent (e.g., race/ethnicity, migration status, language, insurance, informal work, exposure to violence) do not carry the same meaning or risk everywhere. As a result, even strong registries and modern pipelines can fail to support situated equity analysis. This gap is reinforced by three recurring limits in the literature. First, harmonization efforts improve monitoring but can hide structurally patterned missingness and flatten intersectional realities [40,11,37]. Second, digital follow-up systems (e.g., DHIS2, MCTS, regional platforms) remain fragile where exclusion is produced: paper-to-digital transitions, connectivity, workload, and uneven accountability [2,14]. Third, predictive models for maternal risk are growing fast, but studies often under-report behavior under dataset shift and how bias enters through access and documentation [38,25]. These limits motivate our comparative, intersectionality-aware analysis of pregnancy follow-up datasets across Europe, the Americas, Africa, and Asia. We map not only which indicators exist, but also which intersectional dimensions are missing, how categories travel (or fail to travel), and how governance and infrastructure shape the data that AI systems learn from.

3 Comparative landscape and intersectional limits of fetal ultrasound datasets and non-imaging monitoring practices

Public fetal ultrasound datasets have become central benchmarks for computer vision in obstetrics, yet they typically operationalize "pregnancy status" through a small set of standardized imaging targets (e.g., head circumference or abdominal circumference) and a narrow family of tasks (plane classification, segmentation, landmarking, measurement). By contrast, routine antenatal monitoring in real-world care pathways combines imaging with longitudinal, embodied and relational signals—such as serial symphysis–fundal height (SFH), maternal

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

perception of changes in fetal movements, and palpation-based assessment of lie/presentation—that often determine whether, when, and how ultrasound is performed [22,24,36]. This misalignment matters for intersectional and situated analyses: demographic, socio-economic, care-access, and device/operator context variables that shape how pregnancy follow-up is experienced and acted upon are often absent, inconsistently reported, or ethically inaccessible in public dataset releases [12]. As a result, many datasets are well-suited for technical performance comparisons but are poorly equipped to support intersectional analyses of fetal evolution as a socially situated process unfolding over time.

Public ultrasound datasets: regions, pregnancy contexts, and operationalized "status".

A systematic search following PRISMA guidelines [13] was conducted across six repositories (Google Scholar, Scopus, PubMed, Zenodo, PhysioNet, Kaggle) in January 2026 using relevant fetal ultrasound dataset keywords, with no lower date limit.

Records were screened by title/abstract or repository descriptions, and links were checked to confirm dataset availability and content. Datasets were included if publicly accessible, containing 2D/3D fetal ultrasound images with documentation, and excluded if inaccessible, lacking structured downloads, or containing no raw ultrasound images.

Table 1 provides an overview of representative public ultrasound datasets across regions and pregnancy contexts used in this work. It highlights a recurrent pattern: datasets are frequently single-site or concentrated in Europe, while multi-site collections (including those spanning African countries) are small per site and are primarily designed to stress-test generalization across devices and operators rather than to enable inequality analysis [32,33]. Moreover, most releases focus on a narrowly defined measurable endpoint (standard plane recognition, one biometric, or one anatomical segmentation target) so that "status" becomes what is easiest to label rather than what is clinically and socially meaningful to follow longitudinally. For example, HC18 foregrounds head circumference measurement [18,19], while ACOUSLIC-AI emphasizes abdominal circumference estimation from blind sweeps to support low-resource deployments [30,31] in low-income countries (LICs). Intrapartum datasets such as PSFHS further shift the definition of "status" to labor progression, which is not directly comparable to antenatal growth without explicit framing [6]. A different operationalization of "status" appears in the Ultrasound Fetus Dataset [1], which frames fetal condition as a three-class classification problem (normal, benign, malignant); however, this labeling scheme departs from standard obstetric nomenclature, and the absence of provenance metadata, including acquisition site, equipment, patient demographics, and gestational age, makes it impossible to assess population representativeness or to situate the dataset within a broader clinical or sociodemographic context.

Non-imaging monitoring practices and the "missing pathway" problem. Routine pregnancy monitoring includes low-cost, non-imaging practices that are frequently the *gatekeepers* to ultrasound referral and follow-up inten-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

Table 1. Public fetal ultrasound datasets: geography, pregnancy context, and operationalized "status" (targets).

Dataset	Region	Context	Targets / notes on transfer
MFUSPAC [32,33]	Africa (multi-site)	2–3T; device shift	Plane labels; good robustness benchmark; limited harmonized social vars.
FASSD [9]	BR (single site)	Late gestation	Abdomen anatomy segm. (AC-related); late-term/site-specific limits transfer.
FETAL_PLANES_DB [4,5]	ES (Barcelona)	2–3T routine	Standard planes; strong for plane cls.; limited outcomes/socio-demog.
HC18 [18,19]	Netherlands (single site)	1–3T head biometry	HC ellipse/segm.; narrow "status" (HC only); thin context metadata.
ACOUSLIC-AI [30,31]	Sierra Leone	Blind sweeps	AC (frame+segm.+mm); novice/operator generalization; sparse intersectional descriptors.
PSFHS [6]	China (multi-inst.)	Intrapartum	Pubic symphysis+head segm.; labor "status" (not antenatal growth).
UFD [1]	Unknown	Unspecified	3-class cls. (normal/benign/malignant, 1,669 PNG); non-standard obstetric labeling; no provenance metadata.

Note. "Status" denotes what each dataset makes measurable (biometry/anatomy/labor), not a full clinical characterization; intersectional variables are often under-specified [12].

sity. Table 2 highlights three common examples. SFH provides a serial growth trajectory; perceived changes in fetal movements trigger urgent assessment pathways; and Leopold maneuvers inform presentation and engagement late in pregnancy [22,24,36]. These practices are strongly mediated by communication, trust, workload, language, and access to care, and they produce information that is often recorded narratively or not digitized at all. Public imaging datasets rarely include these upstream signals, referral reasons, triage decisions, or longitudinal trajectories. Consequently, what appears as "missing ultrasound data" in a dataset cannot be interpreted: it may reflect a healthy low-risk pathway, but it may also reflect exclusion from care. This ambiguity is especially harmful for equity research because access and triage are socially patterned.

Critical synthesis. Beyond geographic imbalance, three coupled limitations explain why existing public ultrasound datasets rarely support intersectional analyses of fetal evolution. First, they rarely encode pregnancy as a *linked longitudinal process*. Images are commonly released as de-identified snapshots with limited visit-level linkage, restricting analyses of trajectories across gestational weeks and obscuring care discontinuities (missed visits, delayed screening, or uneven follow-up intensity) that are central to equity questions. Second, social descriptors are sparse or non-comparable. Variables needed to examine intersectional inequalities (e.g., structural vulnerability indicators, migration and language barriers, disability, housing insecurity, insurance regime, or locally meaningful categories) are often absent, not standardized, or unavailable due to governance constraints [12]. This prevents auditing whether missingness and measurement error are socially patterned, and it undermines claims about generalization "across populations" when population composition is unknown. Third,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

Table 2. Non-imaging monitoring practices commonly used to track fetal well-being and growth, contrasted with typical dataset operationalization.

Practice (non-imaging)	What it monitors (proxy for "status")	Limitations / what is missing in datasets
Symphysis–fundal height (SFH) [22,27]	Longitudinal growth trajectory screening (serial tape measurements, plotted over time)	Technique- and context-dependent (BMI, protocol). Datasets rarely store SFH trajectories alongside ultrasound frames, limiting multimodal/longitudinal analyses.
Maternal perception of fetal movements [24]	Perceived change in movement patterns used in care pathways	Often narrative (language, stress, work conditions, prior loss) and shaped by access. Public imaging datasets almost never include these lived-context variables.
Abdominal palpation / Leopold maneuvers [36]	Fetal lie, presentation, engagement (late pregnancy emphasis)	Experience-dependent; not easily reducible to a single label; rarely paired with imaging datasets as structured metadata.

Why this matters for intersectionality: these practices are mediated by communication, trust, workload, language, and care access—dimensions that vary across countries and social positions but are typically absent from dataset schemas.

datasets are frequently curated through exclusions that mirror structural inequities: removing low-quality scans, incomplete metadata, high BMI cases, multiples, anomalies, or preterm contexts can erase precisely the pregnancies where barriers to care and documentation are most consequential. Such curation can inflate apparent model performance while reducing relevance for real-world deployment in under-resourced or marginalized settings.

Taken together, these issues show that "data gaps" are not simply missing fields but missing pathways and missing accountability: upstream monitoring practices, triage decisions, and governance conditions that produce pregnancy data are largely absent from dataset schemas. This motivates intersectional, situated auditing across regions that documents not only which imaging targets are available, but also what social context is missing, how categories travel (or fail to travel) across countries, and how longitudinal follow-up is shaped by unequal infrastructures. In the remainder of this paper, we build on this comparative view to analyze publicly available datasets in detail and to articulate design and documentation requirements for pregnancy follow-up datasets that support responsible, equity-aware research.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

4 Mapping the gap: Who Is Missing from the Datasets Representation

The geographical distribution of images across the five datasets considered in this work reveals a stark imbalance in data provenance. As illustrated in Figure 1, the overwhelming majority of available images originates from a single site in Barcelona, Spain, contributed by the Fetal Planes DB dataset alone.

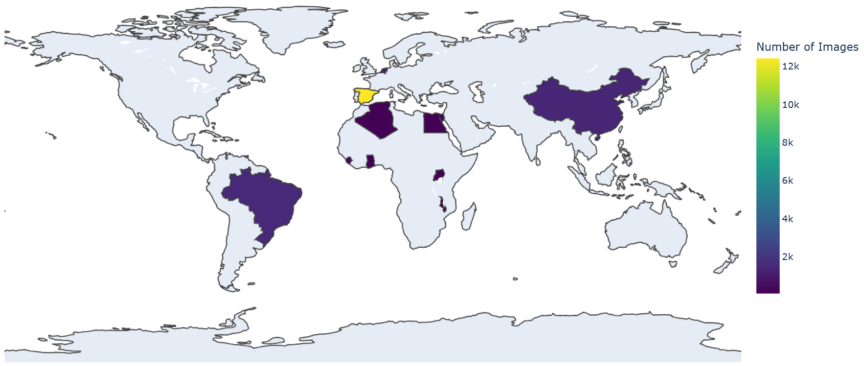

Fig. 1. Geographical distribution of ultrasound images across the five datasets considered in this work.

A secondary, considerably smaller contribution comes from Brazil and the Netherlands, corresponding to the FASSD and HC18 datasets respectively. The African contribution, provided by MFUSPAC across five different countries, is smaller still, representing only a marginal fraction of the total image count. Large portions of the world, including most of the Asian continent, the Middle East, Eastern Europe, and most of the African continent, are entirely absent from the data landscape. However, this geographical skew represents only the most visible layer of a deeper and more insidious form of bias. Even for datasets where the acquisition site is known, no demographic metadata and, in particular, no information on patient ethnicity is provided. This is especially consequential for the Fetal Planes DB, which, despite being by far the largest dataset in terms of image count, offers no information on the ethnic composition of the 1,792 patients enrolled at the two Barcelona hospitals. As a result, it is impossible to determine whether the dataset captures any meaningful degree of ethnic diversity. The same limitation applies to HC18, FASSD, and the UFD, for which demographic information is either absent or entirely undisclosed. Furthermore, even in countries that do appear in the geographical distribution, as could be for Brazil, there is no guarantee of heterogeneous ethnic representativeness. In fact, although the FASSD dataset introduces geographic diversity beyond Europe, its 1,500 images were collected exclusively at a single hospital in Florianópolis,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

Santa Catarina. Given the vast demographic and ethnic heterogeneity of Brazil, spanning continental geographic distances and encompassing diverse populations, a single-centre dataset cannot be considered representative of the national population, let alone of the broader Latin American context. The assumption that data collected in southern Brazil adequately captures the variability present across the country is, at best, an oversimplification, and at worst, a source of systematic bias that remains invisible in the absence of demographic metadata. This argument is further reinforced by the case of the African dataset itself: in fact, the contribution of the MFUSPAC dataset, while commendable in its intent to address this gap, remains insufficient in practice. With only 450 images collected across five African countries, the dataset is too small to provide statistically meaningful coverage of the ethnic and physiological diversity present across the African continent. These critical aspects and the limitations of the datasets currently available in the literature have been systematically compiled in Table 3 below.

Table 3. Key limitations of fetal ultrasound datasets.

Dataset	Data & Size	Key Limitations
MFUSPAC	450 images (25 patients/site); class imbalance across sites.	Marked inter-site quality heterogeneity; no annotation protocol; five different machines.
FASSD	∼1,500 images; uneven images per subject.	Three US machines introduce intensity variability; no inter-rater quantification; single center, full-term only.
Fetal Planes DB	12,400 images; severe class imbalance (Brain 711 vs. Abdomen 4,213).	Variable dimensions across six US systems; single expert annotator; no per-image machine metadata.
HC18	1,334 images; near-duplicates from same sessions.	Variable pixel size (0.052–0.326 mm); single annotator; no demographics or patient count.
ACOUSLIC-AI	Blind sweeps, not curated frames; limited operator stratification.	Inherently variable quality (LIC setting); sparse annotations; incomplete demographic metadata.
PSFHS	Multi-institutional; unbalanced per-center distribution.	No standardized acquisition protocol; intrapartum scope only; cross-site reproducibility unevaluated.
UFD	Severe imbalance: malignant ≈60%; masks halve usable images.	Acquisition protocol not described; annotation methodology undisclosed; *benign/malignant* labels clinically unjustified.

The bias emerging from these data is twofold: ethical and clinical. Fetal ultrasound appearance can vary as a function of maternal body composition, subcutaneous fat distribution, and other physiological characteristics that correlate with ethnicity, all of which affect image quality, acoustic window, and the visual appearance of anatomical structures. A model trained predominantly on images from European patients may therefore fail to generalize to populations with different physiological profiles, not because of architectural limitations, but because the training distribution does not represent the target population. Moreover, as confirmed in the Table 3, the images from African sites exhibit sub-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

stantial heterogeneity in quality, ranging from overexposed acquisitions in Egypt to underexposed and low-sharpness scans in Ghana and Uganda, further limiting their utility for training robust models. In this sense, the mere presence of African data in the corpus, as could be for every country, does not resolve the representativeness problem: data that is too scarce or too noisy to be effectively learned from offers little practical benefit in mitigating bias. The aggregate effect of these limitations is that any model trained on the datasets described in this work risks being, in the most ethically precise sense of the term, Europe-oriented by default, not by deliberate design, but by the structural absence of diverse, high-quality, and demographically characterized data from underrepresented populations. Addressing this gap requires not only the collection of larger and more geographically diverse datasets, but also the systematic inclusion of demographic and ethnic metadata, without which the representativeness of any dataset cannot be meaningfully assessed.

Intersectional scoring. Documenting limitations in narrative form, however accurate, does not support structured cross-dataset comparison. To make these gaps actionable, we propose a lightweight **Intersectional Dataset Scoring Framework** (IDSF) that evaluates each dataset along four dimensions: *Geographic & Site Diversity* (GSD), *Demographic & Ethnic Metadata* (DEM), *Longitudinal Linkage* (LON, i.e., visit-level linkage across gestational time), and *Care-Pathway Context* (CPC).

Specifically, GSD captures the number and diversity of acquisition sites and countries (including LMIC representation); DEM reflects the availability of patient-level variables (e.g., ethnicity, age, BMI); LON assesses whether patient data can be linked across gestational visits; and CPC evaluates the presence of clinical context situating imaging within the antenatal care pathway (e.g., referral reason, SFH trajectory, triage signals).

Each dimension is scored 0 (absent), 1 (partial), or 2 (present and usable), for a maximum of 8. Table 4 reports the results.

Table 4. Intersectional Dataset Scoring Framework (IDSF) applied to public fetal ultrasound datasets. Scores: 0 = absent, 1 = partial, 2 = present. Max = 8.

Dataset	GSD	DEM	LON	CPC	Total
MFUSPAC	2	0	0	0	2/8
FASSD	0	0	1	0	1/8
Fetal Planes DB	0	0	0	0	0/8
HC18	0	0	1	0	1/8
ACOUSLIC-AI	1	0	0	1	2/8
PSFHS	1	0	0	0	1/8
UFD	0	0	0	0	0/8

Scores are assigned through a structured review of each dataset's public documentation, including descriptor papers, repository metadata, and challenge

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

reports. They are assigned objectively, as each dimension encodes the presence, partial availability, or absence of the corresponding information within the dataset.

The scores tell a clear story: no dataset exceeds 2/8, and DEM scores 0 across all datasets, confirming that the absence of socioeconomic and ethnic variables is not a marginal gap but a *structural* feature of the current benchmark ecosystem. The IDSF does not resolve these gaps, but makes them visible and comparable: a necessary precondition for defining minimum documentation standards in obstetric AI.

5 Conclusion and Future Work

This paper reviews public fetal ultrasound datasets with an intersectional, situated lens. We show that benchmarks mostly capture "pregnancy status" through a few imaging tasks (plane detection, one biometric, limited segmentation). They rarely include linked visits across gestation, care-pathway data, or socioeconomic and cultural context. Non-imaging signals that often trigger ultrasound (SFH trends, fetal movement concerns, palpation) are also missing. This is not just a technical gap: it blocks equity audits of missingness, measurement error, and model performance. Overall, fetal monitoring data gaps are gaps in variables, trajectories, pathways, and governance—making responsible obstetric AI harder. Future work should build datasets that support intersectional analysis without harming privacy or sovereignty. Priorities include longitudinal linkage (visits, gestational time, follow-up windows, exclusions), governance-compatible social metadata (structural constraints, with clear limits on what cannot be collected), and encoded care pathways (referrals, triage, key non-imaging alerts) to interpret missing data. Evaluation should focus on real-world transfer, using cross-site/device tests and careful subgroup error analysis. Documentation should also cover ecological and cultural context, such as local norms, trust, and infrastructure limits, to reduce the risk of reproducing inequities.

References

1. Anitha, A.: Ultrasound fetus dataset (2024). https://doi.org/10.17632/yrzzw9m6kk.1, https://doi.org/10.17632/yrzzw9m6kk.1
2. Bhattacharya, A.A., Umar, N., Audu, A., Felix, H., Allen, E., Schellenberg, J., Marchant, T.: Quality of routine facility data for monitoring priority maternal and newborn indicators in dhis2: A case study from gombe state, nigeria. PLOS ONE **14**(1), e0211265 (2019). https://doi.org/10.1371/journal.pone.0211265
3. Bowker, G.C., Star, S.L.: Sorting Things Out: Classification and Its Consequences. MIT Press, Cambridge, MA (1999)
4. Burgos-Artizzu, X.P., et al.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes (2020). https://doi.org/10.1038/s41598-020-67076-5

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

5. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020). https://doi.org/https://doi.org/10.1371/journal.pone.0200412
6. Chen, G., Bai, J., Ou, Z., Lu, Y., Wang, H.: Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head. Scientific data **11**(1), 436 (2024), https://doi.org/10.1038/s41597-024-03266-4
7. Crenshaw, K.: Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum **1989**(1), 139–167 (1989)
8. Crenshaw, K.: Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review **43**(6), 1241–1299 (1991). https://doi.org/10.2307/1229039
9. Da Correggio, K.S., Noya Galluzzo, R., Santos, L.O., Soares Muylaert Barroso, F., Zimmermann Loureiro Chaves, T., Sherlley Casimiro Onofre, A., von Wangenheim, A.: Fetal abdominal structures segmentation dataset using ultrasonic images (2023). https://doi.org/10.17632/4gcpm9dsc3.1
10. D'Ignazio, C., Klein, L.F.: Data Feminism. MIT Press, Cambridge, MA (2020)
11. Euro-Peristat Network: European perinatal health report, 2015–2019 (2022), online report and downloadable data tables available from the Euro-Peristat publications page
12. Fiorentino, M.C., Moccia, S., Cosmo, M.D., Frontoni, E., Giovanola, B., Tiribelli, S.: Uncovering ethical biases in publicly available fetal ultrasound datasets. npj Digital Medicine **8**(1), 355 (2025). https://doi.org/https://doi.org/10.1038/s41746-025-01739-3
13. Fleming, P.S., Koletsi, D., Pandis, N.: Blinded by prisma: are systematic reviewers focusing on prisma and ignoring other guidelines? PLoS One **9**(5), e96407 (2014)
14. Gera, R., Muthusamy, N., Bahulekar, A., Sharma, A., Singh, P., Sekhar, A., Singh, V.: An in-depth assessment of india's mother and child tracking system (MCTS) in rajasthan and uttar pradesh. BMC Health Services Research **15**, 315 (2015). https://doi.org/10.1186/s12913-015-0920-2
15. Giaxi, P., Vivilaki, V., Sarella, A., Harizopoulou, V., Gourounti, K.: Artificial intelligence and machine learning: An updated systematic review of their role in obstetrics and midwifery. Cureus **17**(3), e80394 (Mar 2025). https://doi.org/10.7759/cureus.80394, pMCID: PMC11895402
16. Gitelman, L. (ed.): "Raw Data" Is an Oxymoron. MIT Press, Cambridge, MA (2013)
17. Haraway, D.: Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist Studies **14**(3), 575–599 (1988). https://doi.org/10.2307/3178066
18. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., van Ginneken, B.: Automated measurement of fetal head circumference using 2d ultrasound images (2018). https://doi.org/10.1371/journal.pone.0200412
19. van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. PloS one **13**(8), e0200412 (2018). https://doi.org/10.1371/journal.pone.0200412
20. Langhoff-Roos, J., Krebs, L., Klungsøyr, K., Jakobsson, M., Tapper, A.M., Astolfi, P., Gissler, M.: The nordic medical birth registers—a potential goldmine for clinical research. Acta Obstetricia et Gynecologica Scandinavica **93**(2), 132–137 (2014). https://doi.org/10.1111/aogs.12302

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

21. Moons, K.G., Damen, J.A., Kaul, T., Hooft, L., Navarro, C.A., Dhiman, P., Beam, A.L., Van Calster, B., Celi, L.A., Denaxas, S., et al.: Probast+ ai: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. bmj **388** (2025)
22. guideline NG201, N.: Antenatal care (nice guideline ng201): Recommendations. National Institute for Health and Care Excellence (NICE) (2021), https://www.nice.org.uk/guidance/ng201/chapter/recommendations
23. Nkangu, M., et al.: Mind the data gaps: Comparing the quality of data sources for maternal health services in cameroon. SSM – Health Systems **3**, 100016 (2024). https://doi.org/10.1016/j.ssmhs.2024.100016
24. No, G.t.G.: Reduced fetal movements: Green-top guideline no. 57. Royal College of Obstetricians and Gynaecologists (RCOG) (2011), https://www.rcog.org.uk/media/2gxndsd3/gtg_57.pdf
25. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342
26. Pan American Health Organization: Perinatal information system (SIP) (2024), https://www.paho.org/en/sip, program page describing SIP and access to user/administration manuals
27. Papageorghiou, A.T., Ohuma, E.O., Gravett, M.G., Hirst, J., Da Silveira, M.F., Lambert, A., Carvalho, M., Jaffer, Y.A., Altman, D.G., Noble, J.A., et al.: International standards for symphysis-fundal height based on serial measurements from the fetal growth longitudinal study of the intergrowth-21st project: prospective cohort study in eight countries. bmj **355** (2016). https://doi.org/10.1136/bmj.i5662
28. Philibert, M., Zeitlin, J., Estupiñán-Romero, F., Durox, M., Gissler, M., Euro-Peristat Research Group: Evaluating perinatal health in europe: A comparison of routine population birth data sources. Paediatric and Perinatal Epidemiology (2025). https://doi.org/10.1111/ppe.13178
29. Ruga, T., Zumpano, E., Vocaturo, E., Caroprese, L., Arlia, C.: Bias in dermatological datasets: A critical analysis of the underrepresentation of dark skin tones in melanoma classification images. In: Paszynski, M., Barnard, A.S., Zhang, Y.J. (eds.) Computational Science - ICCS 2025 Workshops - 25th International Conference, Singapore, Singapore, July 7-9, 2025, Proceedings, Part I. Lecture Notes in Computer Science, vol. 15907, pp. 434–448. Springer (2025). https://doi.org/10.1007/978-3-031-97554-7_32
30. Sappia, M.S.: Acouslic-ai: Abdominal circumference operator-agnostic ultrasound measurement in low-income countries using artificial intelligence (2024). https://doi.org/10.5281/zenodo.12697994, data set
31. Sappia, M.S., de Korte, C.L., van Ginneken, B., Ninalga, D., Kondo, S., Kasai, S., Hirasawa, K., Akumu, T., Martín-Isla, C., Lekadir, K., et al.: Acouslic-ai challenge report: Fetal abdominal circumference measurement on blind-sweep ultrasound data from low-income countries. Medical image analysis **105**, 103640 (2025)
32. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ali Ahmed, Y., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L.N., Nalubega Kisembo, H., Goitom Sereke, S., Wanyonyi, S.Z., Temmerman, M., Gratacós, E., Bonet, E., Eixarch, E., Mikolaj, K., Tolsgaard, M.G., Lekadir, K.: Maternal fetal ultrasound planes from low-resource imaging settings in five african countries (2023). https://doi.org/10.5281/zenodo.7540448, data set

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15

33. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ahmed, Y.A., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L.N., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five african countries. Scientific reports **13**(1), 2728 (2023)
34. Shulman, H.B., D'Angelo, D.V., Harrison, L., Smith, R.A., Warner, L.: The pregnancy risk assessment monitoring system (PRAMS): Overview of design and methodology. American Journal of Public Health **108**(10), 1305–1313 (2018). https://doi.org/10.2105/AJPH.2018.304563
35. Simini, F.: Perinatal information system (SIP): a clinical database in latin america and the caribbean. The Lancet **354**(9172), 75 (1999). https://doi.org/10.1016/S0140-6736(05)75345-9
36. Superville, S., Siccardi, M.: Leopold maneuvers (2025), https://www.ncbi.nlm.nih.gov/books/NBK560814/, updated 2024 Jul 27
37. Vargas-Solar, G.: Intersectional study of the gender gap in stem through the identification of missing datasets about women: A multisided problem. Applied Sciences **12**(12), 5813 (2022). https://doi.org/10.3390/app12125813
38. Yang, L., et al.: Prediction models for preterm birth: a systematic review of methods, performance, and risk of bias. Acta Obstetricia et Gynecologica Scandinavica (2023). https://doi.org/10.1111/aogs.14475, systematic review; see DOI for the published version
39. Zeitlin, J., Alexander, S., Barros, H., Blondel, B., Delnord, M., Durox, M., Gissler, M., Macfarlane, A., Euro-Peristat Research Network: Perinatal health monitoring through a european lens: eight lessons from the Euro-Peristat report on 2015 births. BJOG: An International Journal of Obstetrics and Gynaecology **126**(13), 1518–1522 (2019). https://doi.org/10.1111/1471-0528.15857
40. Zeitlin, J., Philibert, M., Estupiñán-Romero, F., Loghi, M., Sakkeus, L., Draušnik, Ž., Recio Alcaide, A., Durox, M., Cap, J., Dimnjakovic, J., Misins, J., Bernal Delgado, E., Thissen, M., Gissler, M., Euro-Peristat Research Group: Developing and testing a protocol using a common data model for federated collection and analysis of national perinatal health indicators in europe. Open Research Europe **3**, 54 (2023). https://doi.org/10.12688/openreseurope.15701.2
41. Zerfu, T.A., Asressie, M., Tareke, A.A., et al.: Contributions of district health information software 2 (DHIS2) to maternal and child health service performance in ethiopia: an interrupted time series mixed-methods study. Archives of Public Health **83**, 173 (2025). https://doi.org/10.1186/s13690-025-01641-0

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_15