Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards equitable AI for women's health:accessible data as a catalyst for innovation

Bianca Schor
Amsterdam University Medical Centers
Kristin Collett Caolo
University of Cambridge
Le Minh Thao Doan
Teesside University
Marta Delfino
Queen Mary University of London
Annalisa Occhipinti
Teesside University
Huiqi Yvonne Lu
University of Oxford
Emma Karoune
ekaroune@turing.ac.uk

The Alan Turing Institute

Article

Keywords: AI, women's health, accessible data, open data, FAIR data, digital health, gender bias, equity, clinical AI

Posted Date: March 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8001150/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Towards equitable AI for women's health: accessible data as a catalyst for innovation

Bianca G. S. Schor[1, 7*†], Kristin Collett Caolo[2, 7†], Le Minh Thao Doan[3, 7†], Marta Delfino[4, 7†], Annalisa Occhipinti[3, 7], Huiqi Yvonne Lu[5, 7], Emma Karoune[6, 7]

[1*]Department of Gynaecology, Amsterdam UMC, The Netherlands.
[2]Department of Sociology, University of Cambridge, UK.
[3]School of Computing, Engineering, and Digital Technologies, Teesside University, UK.
[4]Clinical Pharmacology and Precision Medicine, William Harvey Research Institute, Queen Mary University of London, UK.
[5]Department of Engineering Science, University of Oxford, UK.
[6]Tool, Practices and Systems, The Alan Turing Institute, UK.
[7]AI for Women's Health group, The Alan Turing Institute, UK.

[†]These authors contributed equally to this work.

Abstract

Artificial intelligence (AI) is rapidly advancing across health domains, yet its integration into women's health remains challenged, limited by under-representation in clinical literature and datasets, inconsistent data standards, and a lack of coordinated access to multimodal research-quality data resources. This research maps the current horizon of accessible (i.e. open and accessible on request) data that can contribute to AI development for women's health. Main resources include clinical data repositories, cancer registries, biobanks and published research studies. We summarise data resources related to cancers (breast, cervical, endometrial, and ovarian), chronic and acute health conditions (cardiovascular), under-diagnosed conditions (endometriosis), wearable and vital sign data from remote health monitoring, and discuss other potential resources, such as the broader healthcare data in community care and pharmacy data. We provide a working definition of "women's health", a table centralising key accessible data sources under the level of resources (national registry/clinical study, single/multimodality), and discuss key challenges and opportunities to advance AI research and innovations in the field. To support accessibility and reuse, we also provide

an open-access online repository of curated datasets and offer the wider community the opportunity to add to it. This paper thus offers a cornerstone for building an equitable AI for women's health: it can support future assessments of data completeness, demographic diversity, clinically deployability, methodological benchmarks, licensing, pharmacovigilance, and contributes to highlighting the global AI research in the women's health ecosystem.

Keywords: AI, women's health, accessible data, open data, FAIR data, digital health, gender bias, equity, clinical AI

Corresponding author - Emma Karoune, ekaroune@turing.ac.uk

1 Introduction

Accessing the relevant data is key to improving women's health going forward. Individuals assigned female at birth (AFAB) and people who identify as women have been historically excluded from clinical trials [1–3]. Despite recent efforts to redress this asymmetry – e.g. MESSAGE policy framework in the UK [4]– the resulting gender data gap remains abyssal [5–7]. A recent report 'Closing the women's health gap' by McKinsey estimates closing this gap could potentially boost the global economy by $1 trillion annually by 2040. But more importantly, the report states that, although women live longer, they spend more of their lives in poor health - '*a woman will spend an average of nine years in poor health, affecting her ability to be present and/or productive at home, in the workforce and in the community, and reducing her earning potential'*.

Moreover, the practices and institutions in western medicine remain discriminatory towards certain genders and races, in particular non-white women and transwomen. For example, between 2021-2023 in the UK, Black women were twice more likely to die during pregnancy and after childbirth compared to White women [8]. It is therefore urgent to draw more attention and resources into establishing equitable health datasets (through new data collection or by working to debias existing datasets) and ensuring these data are accessibly archived in line with open and FAIR data best practices.

We argue that the role of open and FAIR data is key to tackling this global health issue. We define open data as "data and content that can be freely used, modified, and shared by anyone for any purpose" [9]. Ideally, data should be made open, however, not all data can be made open due to privacy and legal issues. Therefore, FAIR (Findable, Accessible, Interoperable, Reusable) data [10] can also play an important role in women's health research. FAIR data can for example enable better understanding of available research datasets and ensure sustainable (open or restricted) access for researchers. Ensuring access to large, high-quality, open and/or FAIRly archived datasets can in turn help redress systemic discriminations in healthcare and act as a catalyst to advance knowledge of and boost innovation in women's health.

In this paper, we focus specifically on the growing interest in Artificial Intelligence (AI) as an opportunity to advance women's health [11], but also because the gender data gap raises important ethical and technical challenges in relation to AI for women's health. Indeed, a recent report shows AI research and innovation could help reduce the long-lasting and complex inequalities still present in EU healthcare today [12]. However, there is no widely accepted consensus on what constitutes women's health [13, 14]. We define women's health as any health-related consideration or experience that only or disproportionately affects AFAB individuals, girls, and women and/or affects them differently than men (i.e. the focus of this paper), including the way healthcare systems are providing them with or denying them care. Examples include, but are not limited to: cardiovascular disease; gynaecological complaints and conditions such as premenstrual symptoms (PMS), adenomyosis, and ovarian cancer; autoimmune, neurological, or chronic inflammatory conditions such as fibromyalgia, Alzeihmer's disease, and endometriosis; mental health issues such as postpartum depression or anxiety related to menopause and PMS. We acknowledge this definition remains vague and discuss the ethical considerations associated with it in section 3.

Moreover, AI is rapidly advancing in various healthcare domains, e.g. in radiology [12, 15], but its potential benefits for women's health remain under-explored [16, 17]. AI has the potential to improve patients' outcomes, accelerate drug discovery and help predict conditions [18–20], and yet its integration into women's health remains fragmented and slow [12, 21, 22]. To bridge this gap, there is a pressing need for clinical AI research to explore AFAB and women patient needs, perspectives, voices, and specific challenges in relation to healthcare, due to past and present marginalisation [5, 17]. Moreover, the risks associated with using AI in a high-stake domain like healthcare are high. For example, discriminatory risks towards certain patient populations can directly stem from biases present in the data used for (pre-)training, testing and/or validation. Thus without collecting and accessing the appropriate data, both quantitative and qualitative, it will not be possible to develop clinically-relevant, ethically acceptable, and context-appropriate AI-driven systems for women's health [23].

Our aim with this paper is to facilitate and encourage more research in AI for women's health. Towards this, **we map the current horizon of accessible data relevant to AI development for women's health**. By "accessible data" we mean open access data and/or restricted access data that contains open metadata, which meets or partially meets the FAIR data principles. Indeed, academics raise the access to relevant data as a considerable challenge for their work in this space [14, 21]. Due to our expansive definition of women's health, we do not claim to provide an exhaustive list of relevant data sources. Instead, our review provides a starting point to attract more attention to the field.

Our review spans structured and unstructured sources including biobanks, cancer registries, and data repositories. We examine data resources related to various conditions such as breast, cervical, endometrial, and ovarian cancers; chronic and under-diagnosed conditions such as endometriosis; and broader health signals drawn from wearable devices, primary care records and lifestyle metrics. We exclude private, unpublished datasets, as these are not easily accessible for research purposes, and thus fall outside of the scope of this paper. We include tables of accessible datasets and

highlight key opportunities and bottlenecks in this field. This work therefore offers a foundation for more equitable, transparent, and clinically meaningful AI development in women's health, and calls for the creation of an international women's health open database to list datasets suitable for AI research purposes.

2 Accessible Data in Women's Health: a Fragmented Landscape for AI Research

In this section, we provide an overview of publicly accessible datasets that are relevant for advancing AI research in women's health. AI holds great promise for transforming women's health but its progress is constrained by the quality and availability of relevant data. Drawing on multiple data sources across clinical, imaging, and omics modalities, we highlight opportunities where data availability can support AI innovation, but also major limitations –e.g. gaps in population representation, life-stage coverage, and condition-specific data– that illustrate the fragmented nature of the current data landscape and the barriers it presents for equitable AI development.

2.1 Overview

We review 84 sources of accessible data relevant for AI research in women's health. Table 1 summarises 74 data repositories and datasets and 10 biobanks that meet our inclusion criteria (see section 5). These span various data modalities, from imaging,

Data sources / Data description	**Data repositories**	**Biobanks**	**Total**
Number of accessible data sources identified	74	10	84
including open access sources	47	0	47
including sources relevant for cancer research	24	2	26
Examples of data modalities	imaging, omics, clinical, statistics, histology, cross-modality	clinical, omics, environmental multimodality, activity monitor, measurements and wearables, health data	N/A
Examples of research areas	Biological imaging, cancerology, radiology, gynaecology, cardiology, general health, neurology, epidemiology, maternal health, critical care, ultrasonography, genetics, other diseases	cancerology, rare diseases, genetics, general health, dentistry, maternal health, biology, biochemistry, other conditions	N/A
Geographic areas	UK, EU, USA, Pakistan, India, China, South Africa	UK, EU, USA, China	N/A

Table 1 Overview of accessible data sources relevant to AI for women's health.

omics, clinical histology to transcriptomics, vital signals, and lab results. Among these, we identify at least 47 open source resources, and 26 data sources specifically relevant for AI research in women's cancers. However, we also identify a gap in accessible data that is: i) cross-modal; ii) about non-cancerous conditions, e.g. prevalent gynaecology conditions; iii) representing populations from the global south; iv) about maternal health and menstruation health; and v) combinations of the above-mentioned data. We discuss these limitations further in section 3.

2.2 Accessible Data for AI in Women's Health

Table 2 lists key data repositories and databases we have identified as relevant for AI applications in this field, while Table 3 details the biobanks that provide data resources applicable to AI-driven research in women's health. To guide researchers effectively, we list data sources in alphabetical order with indicative research focus areas, data modality, and type of access (open or request). The following subsections highlight a few key areas of interest for AI research in this mapping of accessible data about women's health. This detailed examination is essential for revealing current data strengths, uncovering critical gaps, and shaping priorities for future data collection. A combined table (Tables 2 and 3) is available on our project's GitHub repository (https://github.com/alan-turing-institute/Data-for-AI-in-Womens-Health). We encourage community contributions and further expansion of our table through opening a GitHub issue or completing a Google form.

Data Repository/Dataset	**Research Focus**	**Modality**	**Link**	**Access**
Abdominal and Direct Fetal ECG Database (ADFECGDB)	Maternal and fetal health	ECG	ADFECGDB	Open
Adolescent Brain Cognitive Development (ABCD Study)	Mental health	Imaging and behavioural data	ABCD Study	Request
BioImage Archive	Biological Imaging	Imaging (microscopy)	BioImage Archive	Open
BRACS: BReAst Carcinoma Subtyping	Breast Cancer	Imaging (H&E histopathological images)	BRACS	Request
Bone Densitometry Dataset	Metabolic and endocrine health	Dual energy x-ray absorptiometry (DXA), Clinical	Bone Densitometry Dataset	Open
Cardiac Atlas Project	Cardiovascular	Imaging (cardiac MRI)	Cardiac Atlas	Open
cBioPortal	Cancer-related	Omics (mutation and RNA-Seq)	cBioPortal	Open
Clinical Practice Research Datalink (CPRD)	Cancer-related and other diseases	Clinical (GP records)	CPRD	Request
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	Cancer-related	Omics (genomics and proteomics) and clinical	CPTAC	Open
Clinical Record Interactive Search (CRIS)	General health, mental health and other diseases	Clinical and others (Statistic)	CRIS	Open
Data Gov	General health and other diseases	Others (Statistic)	Data Gov	Open
DataLoch	General health	Clinical (GP records)	DataLoch	Request

Data Repository/Dataset	Research Focus	Modality	Link	Access
dbGaP	General health and other diseases	Omics (genotype-phenotype)	dbGaP	Request
Dementias Platform UK (DPUK)	Dementia	Omics (genomic), imaging (MRI, PET), Clinical (GP records)	DPUK	Request
Diabetes 130-US Hospitals for Years 1999-2008	Metabolic and endocrine health	Clinical	Diabetes 130-US Hospitals for Years 1999-2008	Open
Discover-NOW	General health	Clinical (GP records)	Discover-NOW	Request
EGA / European Genome-Phenome Archive	Cancer-related and other diseases	Omics (genotype-phenotype)	EGA	Request
European Nucleotide Archive	Cancer-related and other diseases	Omics (genomics and transcriptomics)	ENA	Open
FPUS23	Maternal and fetal health	Imaging (ultrasound images)	FPUS23	Open
GDM Dataset	Maternal health	Clinical	GDM data	Open
GEO datasets	Cancer-related and other diseases	Omics (transcriptomics and genomics)	GEO	Open
GTEx	Cancer-related and other diseases	Omics (transcriptomics, genomics + phenotypes with restricted access), histology	GTEx	Open, Request
HC18 Challenge	Maternal and fetal health	Imaging (ultrasound images)	HC18 data	Open
Heart Disease (UCI Machine learning repository)	Cardiovascular	Clinical (various)	Heart Disease	Open
Heart Failure Clinical Records	Cardiovascular	Clinical (various)	Heart Failure	Open
Health Data Research UK (HDRUK)	General health, cancer-related and other diseases	Clinical, omics (transcriptomics, genomics), imaging (MRI), and others (Statistics)	HDRUK	Request
Healthy and Sustainable Places Data Service	Lifestyle	Lifestyle and others (Statistic)	Data Service	Open
HMBD / Human Metabolome Database	Cancer-related and other diseases	Omics (metabolomics)	HMBD	Open
Human Protein Atlas	Cancer-related and other diseases	Omics (transcriptomics, proteomics)	Protein Atlas	Open
Indian Institute of Science Fetal Heart Sound Database (IIScFHSDB)	Maternal and foetal health	Foetal phonocardiography	IIScFHSDB	Open
Intelligent detection for PCOS datasets	Metabolic and endocrine health	Clinical, imaging (ultrasound), questionnaires	Intelligent detection for PCOS datasets	Open, Request
Irish Hip Fracture Database (IFHD)	Metabolic and endocrine health	Clinical	IFHD	Request
LivWell	Lifestyle	Demographic and socio-economic data	LivWell	Open
Maternal Health and High-Risk Pregnancy Dataset	Maternal health	Clinical	Pregnancy Dataset	Open
Maternal Health Risk Dataset	Maternal health	Clinical	UCI Maternal Dataset	Open

Data Repository/Dataset	Research Focus	Modality	Link	Access
Maternal Mental Health Symptom Profiles Dataset	Maternal mental health	Sociodemographic and clinical data	Mental Health Dataset	Open
MIDRC (NIH, USA)*	COVID-19	Imaging (MRI, X-ray)	MIDRC	Request
MIMIC-III*	Critical care	Clinical (Patient demographics, clinical notes; procedure and diagnosis data (i.e ICD-9 coded data); patient outcome data (length of stay, etc.); imaging reports; physiological (i.e. vital signs, lab reports); medication (i.e. prescriptions and administration details))	MIMIC-III	Request
MIMIC-IV*	Critical care	Same as MIMIC-III	MIMIC-IV	Request
MRC CBU Openly Available Datasets resource	Neurology	Imaging (MRI, fMRI), EEG, behavioural (eye-tracking)	MRC CBU	Open /request
Multi-Modality Ovarian Tumor Ultrasound (MMOTU) image dataset	gynaecology, ultrasonography, cancer-related	Imaging (2D transvaginal ultrasound)	MMOTU	Open
National library of medicine (NCBI)	Cancer-related and other diseases	Omics (transcriptomics, genomics and proteomics)	NCBI	Open
NHS England Digital	Cancer-related, other diseases, gynaecology, maternity	Imaging (MRI, CT, X-ray), clinical (GP records and patient information), others (Statistic)	NHS Digital	Request
Nightingale Open Science	Cancer-related and other diseases	Imaging (X-ray, waveforms, microscopy)	Nightingale	Request
NIMH Data Archive (NDA)	Mental health	Clinical, omics and neurosignal recordings (EEG and EGG)	NDA	Request
OMI-DB	Cancer-related	Imaging (Mammography)	OMI-DB	Request
OpenNeuro	Neurology	Imaging (MRI, PET), EEG	OpenNeuro	Open
OpenSafely	General health	Clinical (GP records, hospital records)	OpenSafely	Request
OrbiGenAI - Maternity Care Modelling	Maternal and fetal health	Demographic, clinical, others (Statistic)	OrbiGenAI	Open
PhysioNet	Cardiology, Neurology	Imaging (X-ray, waveforms), other (multimodal datasets)	PhysioNet	Open
PhysioNet - Noninvasive Fetal ECG	Maternal and fetal health	ECG	Fetal ECG	Open
Pima Indians Diabetes Database	Metabolic and endocrine health	Clinical	Pima Indians Diabetes Database	Open
Pioneer	Cancer-related and other diseases	Clinical (acute care records)	Pioneer	Request
Postnatal maternal depression dataset in South African community cohort	Maternal mental health	Sociodemographic	Postnatal maternal depression	Open
PREdiction of Clinical Outcomes from Genomic profiles	Cancer-related and other diseases	Omics (genomic profiles) and clinical outcomes	PRECOG	Open
Proteomics Identifications Database (PRIDE)	General health, cancer-related and other diseases	Omics (proteomics)	PRIDE	Open

Data Repository/Dataset	Research Focus	Modality	Link	Access
QResearch	Menopause	Clinical and laboratory test data	QResearch	Request
Reproductive Cell Atlas	General health, cancer-related and other diseases	Omics (single-cell omics)	Reproductive Cell Atlas	Open
Roadmap Epigenomics Project	Cancer-related and other diseases	Omics (epigenomics)	Roadmap Epigenomics	Open
SAIL Databank	General health	Clinical and others (statistic)	SAIL Databank	Request
Single Cell Expression Atlas	Cancer-related and other diseases	Omics (single-cell expression data)	Single Cell Atlas	Open
SNAP:EPICCS2*	Critical care	Clinical data	SNAP:EPICCS2	Request
Study of Women's Health Across the Nation	Menopause	Demographics, lifestyle, psychosocial, health outcomes and etc	Swan study	Request
The Cancer Genome Atlas Program (TCGA)	Cancer-related	Imaging (CT, H&E imaging), Omics (transcriptomics, genomics, proteomics) clinical data	TCGA	Open
The Cancer Imaging Archive (TCIA)	Cancer-related	Imaging (CT, PET), Omics (transcriptomics, genomics), clinical data	TCIA	Open
The Health Improvement Network (THIN)	General health	Clinical (electronic health records)	THIN	Request
The Metabolomics Workbench	Maternal health, gynaecology	Omics (metabolomics)	Metabolomics Workbench	Open
The Million Women Study	Lifestyle	Lifestyle	Million Women Study	Request
UNICEF maternal data	Maternal and fetal health	Demographic, Others (Statistic)	UNICEF	Open
USOVA3D Annotated 3D ovarian ultrasound images	gynaecology, ultrasonography	Imaging (3D transvaginal ultrasound)	USOVA3D	Request
UterUS Annotated Dataset of Uteri in Volumetric Ultrasound Data	gynaecology, ultrasonography	Imaging (3D transvaginal ultrasound)	UterUS	Open
WHO	Maternal health	Demographic, Others (Statistic)	WHO	Open
Women's Health Initiative	General health, other diseases	Omics (metabolomics, epigenomics, proteomics, and transcriptomics)	Women's Health Initiative	Request
World Bank Open Data	Lifestyle	Others (Statistic)	World Bank	Open

Table 2: List of publicly available and request-based data repositories relevant to AI research in women's health. Abbreviation List: CT: Computed Tomography, ECG: Electrocardiogram, EEG: Electroencephalogram, fMRI: Functional Magnetic Resonance Imaging, GP: General Practitioner, H&E: Hematoxylin and Eosin, ICD-9: International Classification of Diseases, 9th Revision, MRI: Magnetic Resonance Imaging, PET: Positron Emission Tomography.* indicates datasets derived from crowdsourced sources

2.3 Data Modality

The surveyed data presented in the tables include a wide range of modalities that collectively provide a comprehensive view of women's health. These include clinical data, capturing medical records, diagnostics, and treatment outcomes; omics data,

Biobank name	Modality	Access link	Access	Notes
All of Us	Clinical, omics (genomic), measurements and wearables	researchallofus.org	Request	American population only.
Born in Bradford	Omics, clinical (dental, primary and secondary care, health assessments), environmental	borninbradford.nhs.uk	Request	Bradford based biobank following 60,000 individuals. Includes trios, pregnancy data and dental data.
Breast Cancer Now UK	Omics data	breastcancernow.org	Request	UK-based biobank for breast tissue, breast cells and blood samples from breast cancer patients .
China Kadoorie Biobank	Omics, clinical data	ckbiobank.org	Request	Over 512,000 adult participants. Questionnaires, physical measurements, blood and biological samples
Estonian Biobank	Omics	estonian-biobank	Request	Biobank of Estonia. Mostly genomic array with some WGS and other omics. (There is an application cost on top of compute cost.)
Genomics England	Omics, clinical data	genomicsengland.co.uk	Request	Various datasets: 100K Genomes, NHS Genomic Medicine Services. Both for rare disease and cancer research.
Genes & Health	Omics, health data	genesandhealth.org	Request	Focused on British-Pakistani and -Bangladeshi communities.
Generation Scotland	Omics, clinical data	genscot.ed.ac.uk	Request	Genomics, epigenomics, metabolomics and proteomics of Scottish population ($\tilde{2}$0K individuals).
Our Future Health	Omics (genotype, linked clinical records), health data	ourfuturehealth.org.uk	Request	UK based biobank with a wider range of participant age than UK biobank, and more focus on diverse ethnic backgrounds.
UKBiobank	Omics, other modalities (ECGs, MRIs, activity monitor)	ukbiobank.ac.uk	Request	Can also be used to study women-only conditions or sex-specific features of conditions. $\tilde{2}$73k women, although mostly white-British.

Table 3 List of major biobanks providing valuable data resources for AI research in women's health. The notes column highlights important considerations such as demographic focus, sample sizes, and data types, enabling researchers to identify appropriate resources for their specific AI-driven studies in women's health.

offering molecular-level insights; and medical imaging data, with a particular focus on gynaecological imagery. Furthermore, the growing use of sensors, wearables, and mobile applications has enriched datasets reflecting behavioural, physiological, and

lifestyle dimensions. We now outline the key characteristics, opportunities, and challenges associated with each modality, establishing a foundation for mapping existing datasets and advancing AI-driven research in women's health.

2.3.1 Clinical Informatics Data

Clinical informatics is a clinical information theory and data engineering infrastructure that represents the time-series healthcare information, procedures, and decision making processes, that catalyst digital technologies and data in health aiming to improve healthcare. Data in primary, secondary, and community cares are often stored in different modalities, either raw or processed, but not interlinked. The different data modalities range from patient information, consultation notes, ICD codes, medication codes, to images, videos, and clinical user records.

The differences in data modalities and granularity among unlinked data used for clinical AI make it impossible to achieve generalised intelligence. Thanks to the recent advancement of foundation models and federated learning, unlinked data started to be linked by learning from models, instead of learning from raw medical data. These data are often saved in two forms: structured and unstructured data. Most of AI for health work focuses on structured data and natural human language data. More researchers have started to consider the effect of missing data and/or labels, their effect on information representation, and on AI training and inference learning.

A notable number of sources of clinical informatics data originate from the UK. For example, NHS Digital includes national datasets from care records, systems and organisations on specific areas of health and care [24]. Another source is QResearch, which links GP, cancer registry, civil registration, hospital episode statistics, Intensive Care National Audit and Research Centre (ICNARC), Second Generation Surveillance System (SGSS) COVID test, Pregnancy Registry, and COVID National Immunisation Database (NIMS) data [25]. Other sources include DataLoch [26], with health and social care from Scotland, Discover-NOW [27], and the Clinical Practice Research Datalink (CPRD)[28].

2.3.2 Omics Data

Advances in high-throughput technologies have enabled the comprehensive capture of multiple molecular layers underlying human biology and disease, collectively termed omics data. These datasets, include, but are not limited to, genomic, transcriptomic, epigenomic, proteomic, and metabolomic profiles. While single omic data can provide insight, integrative analysis of such data facilitates a deeper understanding of disease. Such data also supports applications in biomarker discovery and precision medicine, which can be used for both prevention and diagnosis. Multiomics data has shown to be beneficial in understanding, predicting, and modelling multiple women's health diseases, including endometriosis [29], preterm birth [30–32], accelerated biological aging in Asian women [33], and pregnancy [34, 35] . However, while multiomic datasets can provide greater insight and understanding of disease, open datasets remain scarce. In the context of women's health data, we identified biobanks and data repositories as the main sources of omic data, along with one US based longitudinal study [36].

In general, biobanks provide the richest source of multiomic data, as multiple omics can be linked to single individuals at population scale. All biobanks include genomic data, and only two biobanks provide omic data for duos or trios, Genomics England 100K Genomes Study [37], and Born in Bradford [38], with the latter also providing maternal omic data collected during pregnancy. A selection of biobanks also provide other omics, such as transcriptomic and proteomic data (UK Biobank [39], All of Us [40], etc.). It is important to note that all biobanks require an application and varying levels of cost for the application and/or the compute cost required to access data in a secure research environment.

We also identified data repositories for single-omic or multiomic data as another source of women's health omic data. For example, data repository PRIDE [41] contains the proteomic data for Straub's 2025 work on defining lipoedema's molecular hallmarks using multiomics [42]. Another example is the Metabolomics Workbench [43], which holds the metabolomics data from Maric's 2022 paper on multiomics longitudinal modelling of preecplamtic pregnancies [44]. Other omic data from large projects also exist with their own data portals, such as GTEx [45] for transcriptomic or the Human Protein Atlas [46] for proteomic and transcriptomic data. While most data identified was at the tissue or sample level, we only identified one data source for single cell transcriptomics specifically for women's health: the Reproductive Cell Atlas [47]. This source contains single-cell expression data at various stages of development of the reproductive system.

The Women's Health Initiative (WHI) is the largest women's-only longitudinal prevention study conducted in the world, focusing on prevention strategies for heart disease, cancer, and osteoporosis in menopausal women [36]. We found this study is the only longitudinal study that provides access to multiomic data; including genomic, transcriptomic, epigenomic, metabolomic, and proteomic data. However, data is deposited in various locations which all require an application process, with the most sensitive and individual data only available after a longer application process.

2.3.3 Gynaecological Imagery

Imaging is a widespread data modality in gynaecological research and practice and thus yields great opportunities for AI to help advance women's health [16]. This includes mainly MRI, a modality often used in other clinical AI research [12] but onerous; ultrasound scanning, both abdominal or transvaginal, which is rarely standardised but more affordable and accessible; and CT scanning, which combines computer technology with X-ray imaging. Rarer and more invasive imaging also include surgery videos, for example laparoscopy videos are being used for AI research projects and technological innovations [48]. As a reminder, gynaecological imagery includes data about AFAB individuals, i.e. not cis men, but not all gynaecology patients identify as women.

To this day, AI has had a limited impact in gynaecological imagery [16, 48]. Our review lists only three open access datasets containing gynaecological imagery used in AI research: MMOTU [49], UterUS [50], USOVA3D [51]. All of these datasets include transvaginal ultrasound scans exclusively. To our knowledge, there is currently no open data about other gynaecological imagery being used for AI research. These datasets have been published in the last 5 years and, in the same period, we observe an increase

in the number of AI research papers using gynaecological imagery [22]. For example, in the last two years, three literature reviews on AI applied to gynaecological imagery have been published, indicating a growing interest in the field [52, 53]. These trends show that: i) the lack of open gynaecological imagery data is still significant and an important bottleneck for future AI research in gynaecology ii) releasing open gynaecological imaging data is directly translating into more AI research in gynaecology, as shown by Buis et al. [22] iii) although still very limited compared to the main focus of clinical AI –e.g. cardiology, neurology– there is a growing interest in this space. Lastly, the recent literature reviews mentioned above indicate that biobanks are not often used in AI research using gynaecological imagery [22, 52, 53]. This raises questions on whether such data sources effectively enable AI research in relation to gynaecological imagery: are they sufficiently FAIR and known from researchers in this space? Could they be leveraged successfully going forward, for instance in cross-modal AI research projects? We discuss cross-modality further in section 3.2.2.

2.3.4 Data from Sensors, Wearables, and Mobile Apps

Person generated health data (PGHD) are clinically relevant data collected outside of the standard clinical setting. PGHD includes any data collected from wearables such as smart rings, smart watches, fitness trackers, to user-inputted data from apps or websites such as menstrual cycle apps [54]. In the context of women's health, PGHD has been increasing with the development and accessibility of new technologies, especially around menstrual, sexual, and reproductive health [54]. However, we did not see the same trend in open data. To our knowledge, most menstrual cycle data originates from menstrual cycle tracking apps, but is not openly accessible to researchers. For example, the menstrual cycle tracking app Flo (https://flo.health/) accepts requests for collaboration, but application process or timeline is not openly described. On the other hand, the All of Us biobank is the only resource we identified that provides access to PGHD [40]. All of Us includes data from the Fitbit fitness tracker, with around 68 percent of the data from women. Although most of the data are from individuals of white background, they have a program that provides the fitness tracker to increase representation [55].

2.4 Data Resources for Research on Women's Health

Research on women's health encompasses a wide range of topics, including cancer research, cardiovascular health, mental health, menopause, maternal health, and lifestyle and public health. We now highlight key data resources within these research domains, illustrating their potential to support AI applications and guide future data-driven innovation in women's health research.

2.4.1 Cancer Research

Cancer represents one of the most critical areas within women's health where AI has transformative potential, particularly in improving early detection, personalised treatment, and understanding biological heterogeneity. The datasets outlined in Table 2

form a cornerstone of this effort by providing multimodal, high-dimensional data critical for AI model development. Resources such as The Cancer Genome Atlas (TCGA) [56] and The Cancer Imaging Archive (TCIA) [57] offer integrated genomic, transcriptomic, imaging, and clinical datasets that enable researchers to build predictive models for breast, ovarian, cervical, and endometrial cancers. These repositories allow AI systems to link molecular profiles with imaging phenotypes and clinical outcomes, a key step toward precision oncology tailored to women's unique biological and clinical presentations [58].

Parallel to these cancer-centric repositories, the biobanks described in Table 3 provide large-scale population data with diverse omics and clinical variables that are vital for contextualising cancer research in broader health and demographic frames. For example, UK Biobank [39] and Genomics England [59] host extensive genomic and clinical data that include thousands of cancer cases, alongside longitudinal follow-up and health record linkage. These biobanks enable researchers to study sex-specific cancer risk factors, progression patterns, and treatment responses in real-world populations [60, 61]. Notably, the UK Biobank's inclusion of multimodal data such as imaging (MRI, ECGs) and lifestyle measurements allows for richer AI models that can incorporate environmental and physiological variables alongside molecular markers. Additionally, biobanks focusing on underrepresented groups, such as Genes & Health with its British-Pakistani and Bangladeshi cohorts, are crucial to addressing disparities in cancer outcomes by facilitating AI analyses sensitive to genetic and socio-demographic diversity [62].

Despite the availability of rich cancer-focused biobanks and data repositories, significant challenges persist that hinder the development of clinically meaningful and equitable AI tools for women's cancer research. Many cancer datasets remain heavily biased toward populations of European ancestry, limiting representation of minority ethnic groups and rare cancer subtypes that disproportionately affect women, such as triple-negative breast cancer or certain ovarian cancers. This demographic skew not only perpetuates the broader gender data gap but also constrains AI models from capturing the full heterogeneity of women's cancer biology [63]. Moreover, inconsistencies in data harmonisation across repositories, variable quality of metadata, particularly regarding sex, gender, and tumour subtypes, and ethical and legal access restrictions complicate dataset integration and reuse [64, 65]. Addressing these challenges requires ensuring cancer datasets are FAIR, enriched with detailed clinical and demographic metadata, and linked with emerging women's health data streams such as wearable devices and primary care records. Such integrated, high-quality data ecosystems will be essential to developing AI models that accurately reflect both the molecular complexity of women's cancers and their real-world health experiences, thereby improving precision oncology outcomes and reducing health disparities.

2.4.2 Cardiovascular Health

Cardiovascular health is a core focus of women's health. 25,000 women die every day from cardiovascular disease worldwide [66]. While overall cardiovascular mortality has been decreasing for the past 20 years, this decrease is not the same for every age or gender: it is declining for men at any age, but a worrisome increase has been observed in

women under 55 [66]. Disparities in this field are widespread, e.g. in-hospital mortality from acute coronary syndrome is higher for women than men for every age bracket [66]. It is therefore crucial to leverage AI to improve women's health in this field. We observe that cardiovascular data is available for AI research in women's health, the earliest source we have identified being published in 1988 [67]. We have reported three main data sources in Table 2: one imaging source (the Cardiac Atlas Project [68, 69]), and two clinical data sources: the UCI Machine Learning Heart Disease Database that includes the well-known Cleveland dataset [67], and the Heart Failure Clinical Records Dataset, also available in the UCI repository [70]. These sources are open access, with sex included for all patients, and have led to multiple ML research papers [71–74]. These datasets are well referenced and transparent in terms of the data types and data provenance. For example, the Heart Failure Clinical Records Dataset contains clinical data from including female patients from Pakistan – a rare source of data about female population from the Global south. Whilst the datasets are helfpul for AI research in this space, we note that none focuses on female exclusively, therefore the main limitation in this field is the relative small number of female patients included in each dataset. Other relevant and open source datasets exist on Kaggle, however we excluded them from this review as they either include the same data sources as mentioned above, or do not provide information on data provenance, and could potentially include the same data.

2.4.3 Metabolic and Endocrine Health

The metabolic and endocrine health of women is governed by a complex interplay of genetic, environmental, and constitutional factors. Chronic conditions such as osteoporosis and type 2 diabetes demonstrate sex-specific risk factors, disease drivers, and clinical outcomes [75], while others are unique to women, such as polycystic ovary syndrome (PCOS) and gestational diabetes. Reproductive traits across the female lifespan including age at menarche, menstrual irregularity, the development of PCOS, gestational weight change, gestational dysglycemia and dyslipidemia, and the timing and severity of menopausal symptoms, have each been linked to long-term metabolic risk, such as type 2 diabetes [76].

Recent research has successfully developed AI models for various applications in women's metabolic and endocrine conditions. For example, multiple studies have developed AI models for prediction gestational diabetes [77, 78]. For type-2 diabetes, AI has also been used to personalise prescribing [79]. Another study demonstrated the use of imaging from breast scans to predict type-2 diabetes in women [80]. For PCOS, various studies developed AI for prediction, diagnosis, classification, and screening of potential complications. These studies use a variety of data modalities such as imaging, clinical and biochemical markers, biopsy samples, and even data based on tongue and pulse, and full-eye scleral images [81–85].

Looking at PCOS, a review by Li and He et al. identified all current open datasets and concluded that current datasets lack multimodal data and are not maintained [86]. The authors also provide a website listing all the datasets (sites.google.com/view/sokpcos/). Diabetes related datasets were also identified on UC Irvine Machine Learning Repository [87], and on Kaggle [88]. Open datasets for osteoporosis research were

also found [89, 90], along with one surce that requires access, the Irish Hip Fracture database [91].

In general, the vast amounts of multiomic and multimodal data, including clinical and health data, from large biobanks such as UK Biobank [39], All of Us [40], and Born in Bradford [38] have the potential to be used for AI for women's endocrine and metabolic health. Women-only longitudinal studies such as the WHI [36] and the SWAN study [92] also provide longitudinal data on these health conditions.

2.4.4 Mental Health

Mental health represents a critical component of women's health, particularly during vulnerable life stages such as adolescence, pregnancy, postpartum, and menopause. In fact, women are statistically more likely to be affected by common mental health conditions such as depression and anxiety, and are also more prone to trauma-related disorders [93]. Despite this burden, mental health datasets remain underrepresented in AI research for women's health, limiting the development of AI-driven solutions tailored to women's specific needs.

Recent research has begun to explore how AI can support screening, diagnosis, and personalised treatment for mental health conditions in women. For instance, AI models have been developed to predict the onset of postpartum depression using clinical, demographic, and social data [94, 95]. Complementing these efforts, comprehensive national-level data, such as the UK Government's estimates on the prevalence of perinatal mental health conditions, provide essential context on depression, anxiety, and PTSD during and after pregnancy [96]. Similarly, large-scale population studies such as the UK Biobank, which includes mental health questionnaires linked to hospital records, offer opportunities to investigate sex-specific mental health trajectories at scale [97].

Expanding this perspective to earlier developmental stages, resources like the Adolescent Brain Cognitive Development (ABCD) study collect longitudinal brain imaging and behavioural data from over 11,000 adolescents, enabling researchers to analyse early-life mental health and cognitive development with sex-disaggregated insights [98, 99]. In parallel, clinical datasets such as the Clinical Record Interactive Search (CRIS) system provide anonymised real-world mental health records from the UK's NHS Trusts, enabling AI model development grounded in routine care [100]. Additionally, the NIMH Data Archive (NDA) aggregates data from US-funded mental health research projects and supports the development of AI models through its large, diverse datasets across age groups and diagnoses [101].

While these resources are essential for training robust and generalisable AI models, they also offer opportunities to advance more inclusive, equitable approaches to mental health care that account for the unique needs and lived experiences of women.

2.4.5 Menopause

Menopause is an underrepresented but increasingly recognised focus in women's health research, especially in the context of AI. While not many open source datasets are explicitly menopause-specific, several large-scale repositories do capture relevant data

on menopausal status, symptoms, hormone replacement therapy (HRT) use, and associated health outcomes.

For example, the UK Biobank includes self-reported data on age at menopause, use of HRT, and longitudinal health outcomes, allowing for sex-specific and life-stage analyses [97]. Similarly, the QResearch database also captures menopause-related consultations and prescriptions, enabling large-scale observational studies of menopausal symptom management and associated risks [25].

Moreover, cohort studies like SWAN (Study of Women's Health Across the Nation) in the United States provide detailed longitudinal data on hormonal changes, menopausal symptoms, and health outcomes in midlife women, though access is restricted [92]. These datasets offer opportunities to develop AI models that predict symptom severity, optimise HRT strategies, and identify comorbid risks associated with the menopausal transition. However, dedicated, menopause-focused datasets remain scarce, highlighting the need for more targeted data collection initiatives in this space.

2.4.6 Maternal Health

According to WHO, maternal health refers to the health of women during pregnancy, childbirth, and the postnatal period [102]. Despite the improvement in the global health system, maternal morbidity and mortality remain major challenges in many countries. The advancement of AI emerges as a promising tool to enhance maternal health care, reduce mortality risk and access to timely care.

At the global level, the WHO [103] and UNICEF [104] maintain comprehensive datasets that summarise national policies on maternal health alongside socioeconomic indicators such as health expenditure and healthcare coverage. These datasets, together with national-level repositories such as UK Data Gov [105], NHS England [106], and Health Data Research UK (HDRUK) [107], provide the foundation for AI-driven insights that can capture both clinical and social perspectives of maternal outcomes, informing targeted interventions and policy decisions.

Beyond policy-level and socioeconomic data, publicly available data repositories focused specifically on maternal health outcomes have opened new opportunities for AI development in maternal health. Data repositories, such as the UCI Irvine Machine Learning Repository [108] and IEEE [109], provide clinical and demographic information that supports predictive modelling of maternal complications and mortality, facilitating the application of AI in timely interventions for at-risk pregnancies. Other datasets, such as GMD [110] provided retrospective observational data of a large number of pregnancies from EHR, improved delivery planning and maternal monitoring. In addition, datasets investigating maternal mental health are increasingly available, offering opportunities for developing AI models to detect early signs of postnatal depression and anxiety, recommend timely interventions, and improve maternal well-being.

Furthermore, imaging data, particularly ultrasound images, play important roles in maternal and foetal monitoring. Publicly available imaging datasets, such as FPUS23 [111] and HC18 [112], enabled the development of AI models for assessing foetal growth, evaluating placental, and detecting anomalies. In addition to imaging, the

electrocardiogram (ECG) and phonocardiogram also play crucial roles in examining maternal and foetal health during pregnancy. The PhysioNet data repository provides a wide range of maternal ECG and phonocardiogram databases, including synthetic ECG simulators [113] and foetal recording databases (e.g., noninvasive foetal ECG [114], abdominal foetal ECG [115], foetal phonocardiography [116]). These resources facilitate the development of AI models capable of monitoring maternal and foetal health, generating personalised recommendations, and offering predictive insights.

Despite the growing in publicly available datasets for maternal health, many datasets, such as ultrasound imaging and ECG datasets, are restricted and available only upon request due to privacy concerns, which limits the large-scale AI development in this field.

.

2.4.7 Lifestyle and Public Health

Although access to these datasets is typically restricted due to privacy and ethical considerations, several major repositories and platforms provide de-identified or secure access to primary care data for research. These include the Clinical Practice Research Datalink (CPRD), one of the most widely used UK primary care datasets [117], which contains anonymised medical records from over 20 million patients, including diagnoses, prescriptions, referrals, tests, and linked hospital data, and THIN (The Health Improvement Network) [118], a UK-based electronic health record database from general practices, including information on symptoms, diagnoses, prescriptions, and lab results, often used in women's health studies.

In addition to traditional GP datasets, several large-scale, longitudinal repositories have become key resources for investigating sex- and gender-specific trends in primary care. For example, QResearch [25] includes anonymised records from over 35 million patients in general practices, and it is linked to hospital admissions and mortality data, and it has been widely used to model sex-specific disease risks [119]. OpenSAFELY [120], developed in response to the COVID-19 pandemic, allows secure, in situ analysis of electronic health records for over 58 million people in England, supporting granular studies stratified by sex, age, and ethnicity. Similarly, the SAIL Databank (Secure Anonymised Information Linkage) [121], based in Wales, provides access to anonymised health and social care data, including GP records, and has facilitated research into gendered patterns in healthcare access and outcomes.

3 Discussion

We now discuss further some ethical considerations raised by the gender data gap, before highlighting the gaps in accessible data we have identified. Lastly, we describe some key limitations and future research opportunities.

3.1 The Gender Data Gap

In order to reach equitable clinical AI, it is necessary to better contextualise the term 'gender data gap' and its impact on women's health until today. The 'gender data

gap' encompasses the historical exclusion of women from data collection, which in the US stemmed from the exclusion of women from clinical trials [122] resulting in an approximately 30 year history of women's health data available to use in clinical research, technological development, pharmaceutical development and clinical practice. The implications of this gender gap are massive, and without inclusive data collection processes, have led to lack of diversity in data sets, helping to compound issues of inclusion in women's health. We argue that the research community should be aware of the historical context of the gender data gap and its role in driving current data collection initiatives and also the limitations of the current datasets as a result of its short history.

3.2 What is Missing: gaps in accessible data for AI in Women's Health

Based on the data we have mapped, we further define this gender data gap by highlighting different types of data that seem to be currently lacking in accessible data repositories and biobanks. We also mention other potential sources of data relevant to AI research for women's health that could be explored further.

3.2.1 Data about Non-Cancerous Conditions, Mental Health, Maternal Health and Non-White, Non-Cis Populations

As illustrated in Tables 2 and 3, accessible data relevant to AI research in women's health is far from being comprehensive of all women's health conditions and disorders. Table 4 highlights key areas where we think accessible data is missing and urgently needs investing to advance AI in women's health.

Examples of key gaps in accessible data	**Examples of research areas urgent to invest in women's health**
Non-cancerous data	endometriosis*, adenomyosis*, gynaecological laparoscopy data*, transvaginal ultrasound videos*, menstrual cycle data*
Maternal health	maternal health inequalities*, maternal mortality
Mental health	post-partum depression, anxiety related to PMS and (peri-)menaupose*
Non-White and Trans populations	Populations from the Global South*
Cross-modal data	different types of gynaecological imaging combined with clinical data*, combinations of the above-mentioned areas

Table 4 Examples of key gaps in accessible data sources relevant to AI for women's health. * refers to the areas in which no dataset was found from Table 2.

This gap in accessible data is particularly problematic in relation to prevalent conditions considered benign but life-impairing, such as adenomyosis and endometriosis. It dramatically limits potential AI research in this field and the urgent improvement in healthcare that over 190 millions of women need and are waiting for [123]. And yet, our review seems to indicate that no accessible imaging data specifically useful

to these conditions is currently available – despite gynaecological imaging being the first-line diagnostic and management modality used worldwide. Therefore, widespread, non-cancerous disorders and conditions, including chronic ones, such as polycystic ovary syndrome (PCOS), appear to be under-represented in the current accessible data landscape compared to cancerous data.

The gap in accessible data is also glaring in relation to women's mental health, maternal health, and data representing trans individuals and women from the global south. For example, few of the data sources listed in Tables 2 seem to directly relate to women's mental health, despite this area being considered an important concern for future population health going forward. Similarly, only one biobank (Genes & Health) focuses specifically on non-white British communities. Our review also stresses the lack of accessible sources that combines these different types of data, i.e. non-cancerous physical conditions, mental health and/or representing groups from the global south. We further explore cross-modal data below.

Lastly, a notable gap we identified was the lack of accessible person-generated health data, especially menstrual cycle data. A recent review has highlighted the importance of the menstrual cycle as a health indicator, both for gynaecological and general health[124]. However, menstrual cycle patterns are not routinely documented in clinical or research practice, and are therefore not available in accessible clinical data repositories or biobanks. Although menstrual cycle app companies such as Flo accept requests for collaboration with researchers, the process and timelines are unclear. Additionally, there are increasing concerns with such apps, as research shows that there are uncertainties in the privacy and security of the data collected by the applications [125–130]. A recent report by the Minderoo Centre for Technology and Democracy (University of Cambridge) also expressed the need to develop apps owned and operated by public health bodies and to promote stricter regulation [131]. We now focus further on cross-modal data as an opportunity for further AI research.

3.2.2 Cross-Modality Datasets: a Missing Link for AI research

Multimodal or cross-modal datasets, which integrate different types of data such as imaging, genomics, clinical records, and patient-reported outcomes, are increasingly recognised as essential for advancing AI in healthcare [132]. These datasets allow algorithms to learn more complex, nuanced patterns that reflect the biological and social realities of health and disease. In the context of women's health, where conditions like endometriosis or ovarian cancer may present with ambiguous symptoms or complex aetiologies, leveraging multiple modalities can significantly improve diagnostic accuracy, prognostic modelling, and personalised treatment recommendations. Recent studies in gynaecological imaging, for example, show that combining radiology with histopathology or genomic features enhances model performance and interpretability [133, 134]. However, while single-modality datasets are relatively common, fully integrated multimodal datasets remain scarce, especially in publicly accessible repositories.

Tables 2 and 3 report several key resources relevant to women's health that provide, or could potentially be linked across, multiple data modalities. For instance, the

UK Biobank and the All of Us programme both include imaging, genomics, and electronic health records (EHRs), offering a platform for multimodal analysis. Similarly, some cancer-focused datasets, such as TCGA (The Cancer Genome Atlas), provide multi-omic and histopathological data, although they often lack real-world health signals like primary care data or lifestyle information that are critical for women's health contexts. Gynaecological imaging repositories, by contrast, are rarely linked to omics or clinical metadata in an open-access format, which limits opportunities for multimodal research. Integrating such data sources, either through federated frameworks or controlled-access linkages, would help researchers build AI models that are not only more robust but also more contextually aware [65, 135].

Despite the promise, there are considerable challenges to building and using multimodal datasets. These include data governance and privacy issues, inconsistent data standards across modalities, and technical barriers to harmonisation, especially across institutions and health systems [65]. From a gender equity perspective, it is also vital that integrated datasets do not simply aggregate existing biases, such as the overrepresentation of European ancestry or the omission of sex-specific variables, but actively correct for them by ensuring demographic diversity and sex and gender metadata across all modalities. Moreover, the lack of standardised benchmarks and publicly available, well-annotated cross-modal datasets tailored to women's health remains a barrier to innovation. Creating and curating such resources, ideally with FAIR principles in mind, will be critical to unlocking the full potential of AI for understanding and improving women's health across the life course.

3.3 Main Limitations and Future Research

As evidenced in Table 4, key limitations of this paper stem from the limited, varying, and sporadic coverage that the data we have highlighted provide in relation to women's health. Our review confirms the difficulty in finding and/or accessing open and FAIR data relevant to researchers interested in applying AI to women's health.

Firstly, the datasets and repositories listed remain limited and do not represent the breadth of women's health, as defined in the introduction.

Secondly, from a user perspective, we acknowledge it might be easier to review and list the datasets by geographic location, pathology, or data modality; however, our approach seems to indicate that the current landscape of accessible data does not necessarily allow for such systematic classification at this point in time, for there are significant gaps in the accessible data available. For example, we could not find a large, open dataset suitable for AI research that was specifically located in or representing female populations from the global south. Likewise, we observe a strong focus on cancer data and a relative lack of focus on non-cancerous data, as illustrated in the field of gynaecological imagery. While such data might exist and our review is not exhaustive, this is indicative of the difficulty to find such data and the need for our review in centralising accessible data sources. We highlight the data gaps here to encourage more research and accessible data practices to bridge them. We also aim to improve this situation by providing our curated list of data in an open-access online repository to be further edited and expanded by the community (https://github.com/alan-turing-institute/Data-for-AI-in-Womens-Health).

Moreover, through this review, we reveal that there seems to be no consistent standards among the accessible datasets we have identified for women's health research: some datasets have been designed specifically for AI research in women's health (e.g. MMOTU, uterUS) while others might be suitable for AI research but might require domain-specific data cleaning and processing (e.g. metabolomics repository). While beyond the scope of this paper, future research should focus on evaluating AI-readiness of these datasets by assessing completeness, demographic diversity, licensing, and existing methodological benchmarks such as OMOP, CDM, and HL7/FHIR. A framework for assessing gender balance and potential algorithmic impact would also help to make such data more easily usable for AI research in this field, evaluate existing data systematically and create a standard for data that could be used across women's health datasets.

Another important limitation in our paper is the different access requirements for certain datasets, which significantly vary. Therefore, not all sources meet the criteria to be considered "open data". When indicating 'request access,' we acknowledge this term encapsulates various degrees of accessibility, for example, data held in the EU might be easier to access within the EU than from outside the EU for legal reasons. There are also different degrees of FAIRness of these datasets in terms of the metadata that is provided to fully understand and reuse them. In addition, there is variability in cost structure when indicating 'request access' as some datasets are free for the user and some charge a range of fees. Some datasets require training for protecting human research participants (e.g. HIPAA requirements in the US). Others require users to have academic or government affiliations to access and use the data.

Following the open definition, we have defined "open data" according to open licence, open access, open format and machine readability. We thus conclude that too few datasets meet these open data standards in practice. While it goes beyond the scope of this paper to assess to what extent each dataset listed meets these criteria, our review shows that in practice open data remains limited in women's health to this day. We thus call for a shift in practices towards open (data) science from academics, clinicians, and policy-makers involved in the clinical and/or clinical AI space, for example by introducing more incentives and facilitating open data sharing, e.g. with responsible AI/ data science trainings tailored for clinicians collecting data.

If data cannot be made open, we advocate for it to always be made FAIR. Aligning archiving of data with the FAIR data principles does not preclude any privacy or security requirements of the data as it is only the metadata (data about the data) that is made openly available. Using the FAIR data principles ensures that data is sustainability archived in a repository (even in a restricted way) with a persistent identifier. This helps to make it findable, gives the data and metadata a license to allow researchers to understand how it can be used, and ensures the data is standardised for interoperability. Such rich metadata means researchers can fully understand the dataset thus enabling reuse.

Lastly, this paper focuses on human, real data only. We have excluded datasets composed of synthetic and animal data from this review. One example is the International Mouse Phenotyping Consortium data portal, which is a resource of whole gene knock out mouse lines with comprehensive phenotyping data, with an emphasis on

phenotyping both male and female mice ([136]). We acknowledge these might constitute a growing source of open data for AI research on women's health going forward, e.g. due to reduced privacy risks for patients. We thus encourage further efforts to expand our review by including these.

Future research also needs to examine current policy frameworks and methods for the collection of gender and ethnicity data. We also note that in existing datasets it was difficult to find information on gender and ethnicity in both the metadata and the actual datasets, and that these datasets were largely built without these data in mind. This is significant as limited or biased data in AI technology can not only reproduce but amplify bias. Further research into the privacy issues and concerns in women's health data is also necessary.

4 Conclusion

As the role of AI in healthcare is expanding, we show that accessible data can be a catalyst for AI research and innovation in women's health. By reviewing 84 sources of accessible (i.e. open or FAIR) data for women's health, we aim to pave the way for more equitable, transparent, and clinically meaningful AI research in this space. Towards this, we have centralised and analysed 74 data repositories and 10 biobanks, which span various research topics (e.g. cancer, cardiovascular, metabolic and endocrine health, mental health, menopause, maternal health and lifestyle) as well as different data modalities (e.g. omics, imaging, clinical) and geographic locations (e.g. US, Europe, UK, China). We highlight promising opportunities for AI developments, for example, in cancer research where accessible data exists, but also stress important data gaps in particular in i) accessible cross-modal data; ii) non-cancerous conditions; iii) Non-White, non-cis populations; and iv) combinations of the above-mentioned categories (see Table 4). We show that open and FAIR data is still lacking in women's health and that there seems to be no consistent standards among the accessible datasets we have identified, which can limit future AI research but also lead to increased discriminatory risks for female patients. Our main contributions include: i) defining "women's health" and the gender data gap in the field of women's health accessible data, ii) mapping the landscape of accessible data relevant for AI research in women's health, iii) highlighting opportunities, bottlenecks, and future research needed to advance AI developments in women's health. By providing an open-access repository with our list of curated data (https://github.com/alan-turing-institute/Data-for-AI-in-Womens-Health), this review further supports long-term data accessibility and reuse. We thus encourage the communities involved in AI research for women's health to provide additional relevant accessible data sources and policy-makers to fund research and innovative initiatives to bridge the gender data gap and advance women's health with equitable AI.

5 Methods

5.1 Data Collection: Leveraging our Expertise and Network

This mapping stems from various sources: the authors collectively listed all the main datasets that they have used based on their respective expertise in AI research for

women's health and health research more broadly. We then compiled an anonymous questionnaire and shared it online through the UK's national AI for women's health academic group, hosted at the Alan Turing Institute [137]. The survey was sent to individuals researching or working in women's health and included a question that asked if the participant used data in their research; if this data was publicly or privately available; if the data was publicly available, they were asked to specify the database they used or if they were aware of any other public databases in the field; they were then asked what type of data they used. With this community-informed approach, the list of datasets expanded with another four entries, e.g. MIMIC-III, MIMIC-IV, SNAP:EPICCS2, MIDRC.

5.2 Inclusion and Exclusion Criteria

While this overview is not an exhaustive list of data sources, we aimed for consistency and high quality by abiding to the following criteria.

Inclusion criteria:

- human data that fits our definition of women's health (§1) and has been used for AI research (either by the authors or by others)
- only open or FAIR data, following the definitions outlined in §1.

Exclusion criteria:

- synthetic and/or animal data
- data source is unknown/unspecified
- data already available in another, more established source (see example below)
- private data or data only attached to a single paper (i.e. less easily findable or not accessible)

For example, the Cleveland Heart Disease dataset is available both in UCI and Kaggle [67]. We only mention the UCI source here due to its greater data transparency and to avoid repetitions. Each source was checked by at least two authors. Although we do not include it in this paper, we recognize that a large amount of women's health data is collected privately and has been used for women's health research collaborations.

5.3 Online repository

Lastly, we broadened the data collection process by opting for a crowd-sourcing, iterative approach: we designed an online repository of curated datasets (https://github.com/alan-turing-institute/Data-for-AI-in-Womens-Health), where anyone can make suggestions for edits. These will be reviewed by one of the authors on a termly basis before being published.

Declarations

6 Funding Acknowledgements

The work of the AI for Women's Health Group has been funded by The Alan Turing Institute's Clinical AI Interest Group, CRASSH, University of Cambridge and Homerton College, University of Cambridge, UK. HYL would like to acknowledge the support received by the EU Horizons SMARTEdge and AO would like to acknowledge the support received by the EPSRC grant EP/Y001613/1. MD acknowledges support from UK Research and Innovation (UKRI) under the UK Government's Horizon Europe Funding Guarantee (grants 10098097 and 10104323).

7 Competing interests

All authors declare no financial or non-financial competing interests.

8 Ethics approval

We received ethical approval for the survey from the Centre for Research in the Arts, Humanities and Social Sciences (CRASSH) centre at the University of Cambridge, as part of a hybrid event titled "AI in Women's Health: Bridging research and patient voices" organised by the AI for women's health group on 5th June [137]. We confirm our survey methods were carried out in accordance with the relevant guidelines and regulations and informed consent was obtained from all participants.

9 Data availability

The dataset collected for this paper is openly available in Zenodo using this DOI: Schor, B., Caolo, K., Doan, L. M. T., Delfino, M., Occhipinti, A., Lu, H. Y., & Karoune, E. (2025). Towards equitable AI for women's health: accessible data as a catalyst for innovation [Data set]. Zenodo. https://doi.org/10.5281/zenodo.16644069

It is also available to access and add to through our GitHub repository, which can be found here: https://github.com/alan-turing-institute/Data-for-AI-in-Womens-Health

10 Author contribution

All authors have contributed to the conceptualisation, data curation, analysis, investigation, methodology, writing - original draft, and writing - editing and reviewing of this project. BGSS, KCC, LMTD and MD contributed visualisation. BGSS, KCC, AO, EK have contributed to funding aquisition. HYL, EK, AO and BGSS have contributed to the project administration. HYL, AO and EK have provided supervision. All authors have read and agreed to the published version of the manuscript.

References

[1] Vitale, C., Fini, M., Spoletini, I., Lainscak, M., Seferovic, P., Rosano, G.M.: Under-representation of elderly and women in clinical trials. International journal of cardiology **232**, 216–221 (2017)

[2] Shields, K.E., Lyerly, A.D.: Exclusion of pregnant women from industry-sponsored clinical trials. Obstetrics & Gynecology **122**(5), 1077–1081 (2013)

[3] Bennett, J.C.: Inclusion of women in clinical trials–policies for population subgroups. New England Journal of Medicine **329**(4), 288–292 (1993)

[4] MESSAGE: Medical Science Sex and Gender Equity Framework (2025). https://www.messageproject.co.uk/message-policy-framework/

[5] Cleghorn, E.: Unwell Women: A Journey of Medicine and Myth in a Man-made World. Weidenfeld & Nicolson, ??? (2022)

[6] Perez, C.C.: Invisible Women: Data Bias in a World Designed for Men (2019)

[7] Karpel, H.C., Zambrano Guevara, L.M., Rimel, B., Hacker, K.E., Bae-Jump, V., Castellano, T., Curtin, J., Pothuri, B.: The missing data: A review of gender and sex disparities in research. Cancer **131**(6), 35769 (2025)

[8] MBRRACE: Data Brief: Maternal Mortality 2021-2023 (2025). https://www.npeu.ox.ac.uk/mbrrace-uk/data-brief/maternal-mortality-2021-2023

[9] Open Definition: Definition of Open Data (2025). https://opendefinition.org/od/2.1/en/

[10] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., *et al.*: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)

[11] Turco, M.Y., Kraft, O.: A holistic approach to advancing women's health. Nature Reviews Bioengineering, 1–3 (2025)

[12] Directorate-General for Health and Food Safety: Study on the deployment of AI in healthcare (2025). https://health.ec.europa.eu/publications/study-deployment-ai-healthcare-publications-office-eurep_en

[13] Inhorn, M.C.: Defining women's health: a dozen messages from more than 150 ethnographies. Medical anthropology quarterly **20**(3), 345–378 (2006)

[14] Burns, D., Grabowsky, T., Kemble, E., Pérez, L.: Closing the data gaps in women's health. McKinsey & Company (April 2023). https://digitalrosh.com/wp-content/uploads/2023/04/closing-the-data-gaps-in-womens-health-vf-1.

pdf

[15] Andersen, T.O., Nunes, F., Wilcox, L., Coiera, E., Rogers, Y.: Introduction to the Special Issue on Human-Centred AI in Healthcare: Challenges Appearing in the Wild. ACM New York, NY (2023)

[16] Drukker, L., Noble, J., Papageorghiou, A.: Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound in Obstetrics & Gynecology **56**(4), 498–505 (2020)

[17] Schor, B.G.S., Kallina, E., Singh, J., Blackwell, A.: Meaningful transparency for clinicians: Operationalising hcxai research with gynaecologists. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24, pp. 1268–1281. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3630106.3658971 . https://doi.org/10.1145/3630106.3658971

[18] Verma, H., Mlynar, J., Schaer, R., Reichenbach, J., Jreige, M., Prior, J., Evéquoz, F., Depeursinge, A.: Rethinking the role of ai with physicians in oncology: Revealing perspectives from clinical and research workflows. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581506 . https://doi.org/10.1145/3544548.3581506

[19] Bartoletti, I.: Ai in healthcare: Ethical and privacy challenges. In: Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17, pp. 7–10 (2019). Springer

[20] Shah, R., Chircu, A.: Iot and ai in healthcare: A systematic literature review. Issues in Information Systems **19**(3) (2018)

[21] Tank, D., Schor, B.G.S., Trommelen, L., Huirne, J.A.F., Calixto, I., De Leeuw, R.A.: Automatic uterus segmentation in transvaginal ultrasound using u-net and nnu-net. PLOS ONE (2025)

[22] Buis, E.M., Tank, D., De Leeuw, R.A., Trommelen, L., Huirne, J.A.F., Calixto, I., Schor, B.G.S.: Automatic segmentation in transvaginal ultrasound: A systematic review from a computer science approach. Heliyon (2025)

[23] Schor, B.: Designing meaningful algorithmic system transparency for non-expert users. PhD thesis, University of Cambridge (2024)

[24] NHS Digital: Data sets. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets. Accessed: 24 October 2025 (2025). https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets

[25] Hippisley-Cox, J., Stables, D., Pringle, M.: Qresearch: a new general practice database for research. Informatics in primary care **12**(1), 49–50 (2004)

[26] DataLoch: Data – DataLoch. https://dataloch.org/data. Accessed: 24 October 2025 (2025). https://dataloch.org/data

[27] Discover-NOW: Discover-NOW. https://discover-now.co.uk/. Accessed: 24 October 2025 (2025). https://discover-now.co.uk/

[28] Datalink, C.P.R.: Home. https://www.cprd.com/. Accessed: 24 October 2025 (2025). https://www.cprd.com/

[29] Akter, S., Xu, D., Nagel, S.C., Bromfield, J.J., Pelch, K., Wilshire, G.B., Joshi, T.: Machine learning classifiers for endometriosis using transcriptomics and methylomics data. Frontiers in genetics **10**, 766 (2019)

[30] Espinosa, C.A., Khan, W., Khanam, R., Das, S., Khalid, J., Pervin, J., Kasaro, M.P., Contrepois, K., Chang, A.L., Aghaeepour, N., al.: Multiomic signals associated with maternal epidemiological factors contributing to preterm birth in low- and middle-income countries. Science Advances **9**(21) (2023) https://doi.org/10.1126/sciadv.ade7692

[31] Jehan, F., Sazawal, S., Baqui, A.H., Nisar, M.I., Dhingra, U., Khanam, R., Ilyas, M., Dutta, A., Mitra, D.K., Mehmood, U., *et al.*: Multiomics characterization of preterm birth in low-and middle-income countries. JAMA Network Open **3**(12), 2029655–2029655 (2020)

[32] Huang, C., Gin, C., Fettweis, J., Foxman, B., Gelaye, B., MacIntyre, D.A., Subramaniam, A., Fraser, W., Tabatabaei, N., Callahan, B.: Meta-analysis reveals the vaginal microbiome is a better predictor of earlier than later preterm birth. BMC biology **21**(1), 199 (2023)

[33] Chen, L., Tan, K.M.-L., Xu, J., Mishra, P., Mir, S.A., Gong, M., Narasimhan, K., Ng, B., Lai, J.S., Tint, M.T., *et al.*: Exploring multi-omics and clinical characteristics linked to accelerated biological aging in asian women of reproductive age: insights from the s-presto study. Genome Medicine **16**(1), 128 (2024)

[34] Kharb, S., Joshi, A.: Multi-omics and machine learning for the prevention and management of female reproductive health. Frontiers in Endocrinology **14**, 1081667 (2023)

[35] Ghaemi, M.S., DiGiulio, D.B., Contrepois, K., Callahan, B., Ngo, T.T., Lee-McMullen, B., Lehallier, B., Robaczewska, A., Mcilwain, D., Rosenberg-Hasson, Y., *et al.*: Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. Bioinformatics **35**(1), 95–103 (2019)

[36] Study, T.W.H.I., *et al.*: Design of the women's health initiative clinical trial and observational study. Controlled Clinical Trials **19**(1), 61–109 (1998)

[37] 100, .G.P.P.I.: 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. New England Journal of Medicine **385**(20), 1868–1880 (2021)

[38] Wright, J., Small, N., Raynor, P., Tuffnell, D., Bhopal, R., Cameron, N., Fairley, L., Lawlor, D.A., Parslow, R., Petherick, E.S., *et al.*: Cohort profile: the born in bradford multi-ethnic family cohort study. International journal of epidemiology **42**(4), 978–991 (2013)

[39] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.*: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)

[40] Us Research Program Investigators, A.: The "all of us" research program. New England Journal of Medicine **381**(7), 668–676 (2019)

[41] Perez-Riverol, Y., Bandla, C., Kundu, D.J., Kamatchinathan, S., Bai, J., Hewapathirana, S., John, N.S., Prakash, A., Walzer, M., Wang, S., *et al.*: The pride database at 20 years: 2025 update. Nucleic acids research **53**(D1), 543–553 (2025)

[42] Straub, L.G., Funcke, J.-B., Joffin, N., Joung, C., Al-Ghadban, S., Zhao, S., Zhu, Q., Kruglikov, I.L., Zhu, Y., Langlais, P.R., *et al.*: Defining lipedema's molecular hallmarks by multi-omics approach for disease prediction in women. Metabolism **168**, 156191 (2025)

[43] Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K.S., *et al.*: Metabolomics workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic acids research **44**(D1), 463–470 (2016)

[44] Marić, I., Contrepois, K., Moufarrej, M.N., Stelzer, I.A., Feyaerts, D., Han, X., Tang, A., Stanley, N., Wong, R.J., Traber, G.M., et al.: Early prediction and longitudinal modeling of preeclampsia from multiomics. Patterns **3**(12) (2022)

[45] Consortium, G.: The gtex consortium atlas of genetic regulatory effects across human tissues. Science **369**(6509), 1318–1330 (2020)

[46] Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., *et al.*: Tissue-based map of the human proteome. Science **347**(6220), 1260419 (2015)

[47] Marečková, M., Massalha, H., Lorenzi, V., Vento-Tormo, R.: Mapping human

reproduction with single-cell genomics. Annual Review of Genomics and Human Genetics **23**, 523–547 (2022)

[48] Guo, K., Tao, H., Zhu, Y., Li, B., Fang, C., Qian, Y., Yang, J.: Current applications of artificial intelligence-based computer vision in laparoscopic surgery. Laparoscopic, Endoscopic and Robotic Surgery **6**(3), 91–96 (2023)

[49] Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., Chen, L.: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. CoRR (2022)

[50] Boneš, E., Gergolet, M., Bohak, C., Lesar, Marolt, M.: Automatic Segmentation and Alignment of Uterine Shapes from 3D Ultrasound Data. Computers in Biology and Medicine **178**, 108794 (2024) https://doi.org/10.1016/j.compbiomed.2024.108794

[51] Potočnik, B., Munda, J., Reljič, M., Rakić, K., Knez, J., Vlaisavljević, V., Sedej, G., Cigale, B., Holobar, A., Zazula, D.: Public database for validation of follicle detection algorithms on 3d ultrasound images of ovaries. Computer Methods and Programs in Biomedicine **196**, 105621 (2020)

[52] Moro, F., Ciancia, M., Zace, D., Vagni, M., Tran, H.E., Giudice, M.T., Zoccoli, S.G., Mascilini, F., Ciccarone, F., Boldrini, L., *et al.*: Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review. International journal of cancer **155**(10), 1832–1845 (2024)

[53] Moro, F., Giudice, M., Ciancia, M., Zace, D., Baldassari, G., Vagni, M., Tran, H., Scambia, G., Testa, A.: Application of artificial intelligence to ultrasound imaging for benign gynecological disorders: systematic review. Ultrasound in Obstetrics & Gynecology **65**(3), 295–302 (2025)

[54] Karim, J.L., Wan, R., Tabet, R.S., Chiu, D.S., Talhouk, A.: Person-generated health data in women's health: Scoping review. Journal of medical Internet research **26**, 53327 (2024)

[55] Bailey, C.P., Dodd, K.W., McClain, J.J., Seo, I., Wheeler, W., Wolff-Hughes, D.L.: Fitbit physical activity and sleep data in the all of us research program: Data exploration and processing considerations for research. Medicine and science in sports and exercise, 10–1249 (2025)

[56] Network, C.G.A., *et al.*: Comprehensive molecular characterization of human colon and rectal cancer. Nature **487**(7407), 330 (2012)

[57] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., *et al.*: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging **26**(6), 1045–1057 (2013)

[58] Fountzilas, E., Pearce, T., Baysal, M.A., Chakraborty, A., Tsimberidou, A.M.: Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digital Medicine **8**(1), 75 (2025)

[59] Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., et al.: National Genomic Research Library. https://doi.org/10.6084/m9.figshare.4530893.v7 . https://doi.org/10.6084/m9.figshare.4530893.v7

[60] Khan, M., Papier, K., Pirie, K.L., Key, T.J., Atkins, J., Travis, R.C.: Sex differences in cancer incidence: Prospective analyses in the uk biobank. British Journal of Cancer, 1–11 (2025)

[61] Gallagher, C.S., Ginsburg, G.S., Musick, A.: Biobanking with genetics shapes precision medicine and global health. Nature Reviews Genetics **26**(3), 191–202 (2025)

[62] Lee, Y.S., Garrido, N.L.B., Lord, G., Maggio, Z.A., Khomtchouk, B.B.: Ethical considerations for biobanks serving underrepresented populations. Bioethics **39**(3), 240–249 (2025)

[63] Lau, P.L.: Ai gender biases in women's healthcare: Perspectives from the united kingdom and the european legal space. In: YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of Artificial Intelligence, pp. 247–274. Springer, ??? (2023)

[64] Cheng, C., Messerschmidt, L., Bravo, I., Waldbauer, M., Bhavikatti, R., Schenk, C., Grujic, V., Model, T., Kubinec, R., Barceló, J.: A general primer for data harmonization. Scientific data **11**(1), 152 (2024)

[65] Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature medicine **28**(9), 1773–1784 (2022)

[66] Foundation, W.C.H.: Why we're on Red Alert for Cardiovascular Disease in Women (2025). https://www.womencardiovascularhealthcarefoundation.com/alerter/maladies-cardio-vasculaires/Maladies-cardio-vasculaires-des-femmes-alerte-rouge

[67] repository, U.M.: Heart Disease (1988). https://archive.ics.uci.edu/dataset/45/heart+disease

[68] NIH: Cardiac Atlas (2011). https://www.cardiacatlas.org/

[69] Fonseca, C.G., Backhaus, M., Bluemke, D.A., Britten, R.D., Chung, J.D., Cowan, B.R., Dinov, I.D., Finn, J.P., Hunter, P.J., Kadish, A.H., *et al.*: The cardiac atlas project—an imaging database for computational modeling and statistical atlases of the heart. Bioinformatics **27**(16), 2288–2295 (2011)

[70] Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M., Raza, M.A.: Survival analysis of heart failure patients: A case study. PloS one **12**(7), 0181001 (2017)

[71] He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., Zhang, K.: The practical implementation of artificial intelligence technologies in medicine. Nature medicine **25**(1), 30–36 (2019)

[72] Dave, D., Naik, H., Singhal, S., Patel, P.: Explainable ai meets healthcare: A study on heart disease dataset. arXiv preprint arXiv:2011.03195 (2020)

[73] Srinivasan, S., Gunasekaran, S., Mathivanan, S.K., M. B, B.A.M., Jayagopal, P., Dalu, G.T.: An active learning machine technique based prediction of cardiovascular heart disease from uci-repository database. Scientific reports **13**(1), 13588 (2023)

[74] Dawes, T.J., Marvao, A., Shi, W., Fletcher, T., Watson, G.M., Wharton, J., Rhodes, C.J., Howard, L.S., Gibbs, J.S.R., Rueckert, D., *et al.*: Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac mr imaging study. Radiology **283**(2), 381–390 (2017)

[75] Sriram, U.: Applying a women's health lens to endocrine and metabolic disorders. Indian Journal of Endocrinology and Metabolism **25**(3), 171–175 (2021)

[76] Nichols, A.R., Chavarro, J.E., Oken, E.: Reproductive risk factors across the female lifecourse and later metabolic health. Cell metabolism **36**(2), 240–262 (2024)

[77] Kokori, E., Olatunji, G., Aderinto, N., Muogbo, I., Ogieuhi, I.J., Isarinade, D., Ukoaka, B., Akinmeji, A., Ajayi, I., Chidiogo, E., *et al.*: The role of machine learning algorithms in detection of gestational diabetes; a narrative review of current evidence. Clinical Diabetes and Endocrinology **10**(1), 18 (2024)

[78] Zhang, Z., Yang, L., Han, W., Wu, Y., Zhang, L., Gao, C., Jiang, K., Liu, Y., Wu, H.: Machine learning prediction models for gestational diabetes mellitus: meta-analysis. Journal of medical Internet research **24**(3), 26634 (2022)

[79] Dennis, J.M., Young, K.G., Cardoso, P., Güdemann, L.M., McGovern, A.P., Farmer, A., Holman, R.R., Sattar, N., McKinley, T.J., Pearson, E.R., *et al.*: A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study. The Lancet **405**(10480), 701–714 (2025)

[80] Yashar, M.M., Izci, I.B., Gungoren, F.Z., Eren, A.A., Mert, A.A., Durur-Subasi, I.I.: Can artificial intelligence detect type 2 diabetes in women by evaluating the pectoral muscle on tomosynthesis: diagnostic study. Insights into Imaging

15(1), 99 (2024)

[81] Barrera, F.J., Brown, E.D., Rojo, A., Obeso, J., Plata, H., Lincango, E.P., Terry, N., Rodríguez-Gutiérrez, R., Hall, J.E., Shekhar, S.: Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. Frontiers in endocrinology **14**, 1106625 (2023)

[82] Verma, P., Maan, P., Gautam, R., Arora, T.: Unveiling the role of artificial intelligence (ai) in polycystic ovary syndrome (pcos) diagnosis: A comprehensive review. Reproductive Sciences **31**(10), 2901–2915 (2024)

[83] Wang, J., Chen, R., Long, H., He, J., Tang, M., Su, M., Deng, R., Chen, Y., Ni, R., Zhao, S., et al.: Artificial intelligence in polycystic ovarian syndrome management: past, present, and future. La radiologia medica, 1–33 (2025)

[84] Lv, W., Song, Y., Fu, R., Lin, X., Su, Y., Jin, X., Yang, H., Shan, X., Du, W., Huang, Q., *et al.*: Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images. Frontiers in Endocrinology **12**, 789878 (2022)

[85] Wang, W., Zeng, W., He, S., Shi, Y., Chen, X., Tu, L., Yang, B., Xu, J., Yin, X.: A new model for predicting the occurrence of polycystic ovary syndrome: Based on data of tongue and pulse. Digital health **9**, 20552076231160323 (2023)

[86] Li, M., He, Z., Shi, L., Lin, M., Li, M., Cheng, Y., Liu, H., Xue, L., Said, K.S., Yusuf, M., et al.: Intelligent detection for polycystic ovary syndrome (pcos): Taxonomy, datasets and detection tools. Computational and Structural Biotechnology Journal (2025)

[87] Clore, J., Cios, K., DeShazo, J., Strack, B.: Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5230J (2014)

[88] Repository, U.M.L., Kaggle: Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data. Accessed: 2025-10-07 (n.d.)

[89] Masnabadi, N., Sadeghi-Niaraki, A., Karimi, M., AbuHmed, T., Azarbani, N., Choi, S.-M.: Bone densitometry dataset for computer aided osteoporosis disease detection. medRxiv, 2025–01 (2025)

[90] Wani, I.M., Arora, S.: Knee X-ray Osteoporosis Database (Version 2). https://data.mendeley.com/datasets/fxjm8fb6mw/2. Accessed: 2025-10-07 (2021). https://doi.org/10.17632/fxjm8fb6mw.2

[91] Asri, N.A.M., Muslim, A., O'Regan, N., McEliggott, A.: Bone-a-fide breakthrough: Machine learning cracks the code on osteoporosis treatment using the irish hip fracture database. Age and Ageing **53**(Supplement_4), 178–027 (2024) https://doi.org/10.1093/ageing/afae178.027

[92] El Khoudary, S.R., Greendale, G., Crawford, S.L., Avis, N.E., Brooks, M.M., Thurston, R.C., Karvonen-Gutierrez, C., Waetjen, L.E., Matthews, K.: The menopause transition and women's health at midlife: a progress report from the study of women's health across the nation (swan). Menopause **26**(10), 1213–1227 (2019)

[93] Li, S.H., Graham, B.M.: Why are women so vulnerable to anxiety, trauma-related and stress-related disorders? the potential role of sex hormones. The Lancet Psychiatry **4**(1), 73–82 (2017)

[94] Cellini, P., Pigoni, A., Delvecchio, G., Moltrasio, C., Brambilla, P.: Machine learning in the prediction of postpartum depression: A review. Journal of Affective Disorders **309**, 350–357 (2022)

[95] Xie, Y., Zheng, H., Gan, W., Su, C., Shams, M., Yang, J.: The performance of machine learning models in predicting postpartum depression: a meta-analysis and systematic review. Journal of Reproductive and Infant Psychology, 1–18 (2025)

[96] Public Health England: Estimated prevalence of perinatal mental health conditions: England, 2016 to 2019 (2020). https://www.gov.uk/government/publications/perinatal-mental-health-condition-prevalence/estimated-prevalence-of-perinatal-mental-health-conditions-in-england-2016-to-2019

[97] Davis, K.A., Coleman, J.R., Adams, M., Allen, N., Breen, G., Cullen, B., Dickens, C., Fox, E., Graham, N., Holliday, J., *et al.*: Mental health in uk biobank–development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. BJPsych open **6**(2), 18 (2020)

[98] Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., *et al.*: The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. Developmental cognitive neuroscience **32**, 43–54 (2018)

[99] Jernigan, T.L., Brown, S.A., Dowling, G.J.: The adolescent brain cognitive development study. Journal of research on adolescence: the official journal of the Society for Research on Adolescence **28**(1), 154 (2018)

[100] Chilman, N., Song, X., Roberts, A., Tolani, E., Stewart, R., Chui, Z., Birnie, K., Harber-Aschan, L., Gazard, B., Chandran, D., *et al.*: Text mining occupations from the mental health electronic health record: a natural language processing

approach using records from the clinical record interactive search (cris) platform in south london, uk. BMJ open **11**(3), 042274 (2021)

[101] Mental Health (NIMH), N.I.: NIMH Data Archive (NDA) (2024). https://www.nimh.nih.gov/

[102] WHO: Maternal health. https://www.who.int/health-topics/maternal-health#tab=tab_1

[103] WHO: Maternal, newborn, child and adolescent health and ageing - Data portal. https://platform.who.int/data/maternal-newborn-child-adolescent-ageing

[104] UNICEF: Maternal health data - UNICEF DATA (2025). https://data.unicef.org/resources/dataset/maternal-health-data/

[105] data.gov.uk: Find open data - data.gov.uk (2025). https://www.data.gov.uk/

[106] England, N.: Data sets - NHS England Digital (2025). https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets

[107] UK, H.D.R.: Health Data Research Gateway (2025). https://healthdatagateway.org/en

[108] Ahmed, M.: Maternal Health Risk. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DP5D (2020)

[109] Chayan, A.R.: Maternal Health and High-Risk Pregnancy Dataset. https://doi.org/10.21227/ddfa-mf77 . https://dx.doi.org/10.21227/ddfa-mf77

[110] AlSaad, R., Farrell, T., Cruz, J.D., Abd-Alrazaq, A., Thomas, R., Sheikh, J.: Artificial intelligence models for predicting the mode of delivery in maternal care. Journal of Gynecology Obstetrics and Human Reproduction, 102976 (2025)

[111] Prabakaran, B.S., Hamelmann, P., Ostrowski, E., Shafique, M.: Fpus23: an ultrasound fetus phantom dataset with deep neural network evaluations for fetus orientations, fetal planes, and anatomical features. IEEE Access **11**, 58308–58317 (2023)

[112] Heuvel, T.L., Bruijn, D., Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. PloS one **13**(8), 0200412 (2018)

[113] Andreotti, F., Behar, J., Zaunseder, S., Oster, J., Clifford, G.D.: An open-source framework for stress-testing non-invasive foetal ecg extraction algorithms. Physiological measurement **37**(5), 627 (2016)

[114] Behar, J.A., Bonnemains, L., Shulgin, V., Oster, J., Ostras, O., Lakhno, I.: Non-invasive fetal electrocardiography for the detection of fetal arrhythmias. Prenatal

diagnosis **39**(3), 178–187 (2019)

[115] Jezewski, J., Matonia, A., Kupka, T., Roj, D., Czabanski, R.: Determination of fetal heart rate from abdominal signals: evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram. Biomedizinische Technik/Biomedical Engineering **57**(5), 383–394 (2012)

[116] Bhaskaran, A., Kumar, S., George, S., Arora, M.: Heart rate estimation and validation algorithm for fetal phonocardiography. Physiological Measurement **43**(7), 075008 (2022)

[117] Herrett, E., Gallagher, A.M., Bhaskaran, K., Forbes, H., Mathur, R., Van Staa, T., Smeeth, L.: Data resource profile: clinical practice research datalink (cprd). International journal of epidemiology **44**(3), 827–836 (2015)

[118] Blak, B.T., Thompson, M., Dattani, H., Bourke, A.: Generalisability of the health improvement network (thin) database: demographics, chronic disease prevalence and mortality rates. Informatics in primary care **19**(4) (2011)

[119] Snelling, A.J., Copland, E., Mei, W., Mtika, W.M., Ranger, T., Coupland, C., Sheikh, A., Hippisley-Cox, J., Hirst, J.A., Consortium, Q.P., et al.: Establishing the qresearch pregnancy register in the qresearch® database: methods and cohort description (2024)

[120] Nab, L., Schaffer, A.L., Hulme, W., DeVito, N.J., Dillingham, I., Wiedemann, M., Andrews, C.D., Curtis, H., Fisher, L., Green, A., *et al.*: Opensafely: A platform for analysing electronic health records designed for reproducible research. Pharmacoepidemiology and drug safety **33**(6), 5815 (2024)

[121] Ford, D.V., Jones, K.H., Verplancke, J.-P., Lyons, R.A., John, G., Brown, G., Brooks, C.J., Thompson, S., Bodger, O., Couch, T., *et al.*: The sail databank: building a national architecture for e-health research and evaluation. BMC health services research **9**(1), 157 (2009)

[122] Pardue, M.-L., Wizemann, T.M.: Exploring the biological contributions to human health: does sex matter? (2001)

[123] Becker, C.M., Bokor, A., Heikinheimo, O., Horne, A., Jansen, F., Kiesel, L., King, K., Kvaskoff, M., Nap, A., Petersen, K., *et al.*: Eshre guideline: endometriosis. Human reproduction open **2022**(2), 009 (2022)

[124] Vollmar, A.K.R., Mahalingaiah, S., Jukic, A.M.: The menstrual cycle as a vital sign: a comprehensive review. F&S Reviews **6**(1), 100081 (2025)

[125] Amelang, K.: In: Hepp, A., Jarke, J., Kramp, L. (eds.) (Not) Safe to Use: Insecurities in Everyday Data Practices with Period-Tracking Apps, pp. 297–321. Springer, Cham (2022)

[126] Zadushlivy, N., Biviji, R., Williams, K.S.: Exploration of reproductive health apps' data privacy policies and the risks posed to users: qualitative content analysis. Journal of Medical Internet Research **27**, 51517 (2025)

[127] Punzi, M.C., Thuis, T.: Mapping ethical concerns in algorithm-driven period and fertility tracking technologies. Contraception, 110837 (2025)

[128] Hammond, E., Burdon, M.: Intimate harms and menstrual cycle tracking apps. Computer Law & Security Review **55**, 106038 (2024)

[129] Kemp, K.: Your body, our data: Unfair and unsafe privacy practices of popular fertility apps. UNSW Law Research (2023)

[130] Mohan, S., Jenkins, J.: Flowing data: women's views and experiences on privacy and data security when using menstrual cycle tracking apps. Oxford Open Digital Health **3**, 011 (2025)

[131] Felsberger, S.: The high stakes of tracking menstruation (2025)

[132] Occhipinti, A., Verma, S., Angione, C., *et al.*: Mechanism-aware and multimodal ai: beyond model-agnostic interpretation. Trends in Cell Biology **34**(2), 85–89 (2024)

[133] Boehm, K.M., Aherne, E.A., Ellenson, L., Nikolovski, I., Alghamdi, M., Vázquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., *et al.*: Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer **3**(6), 723–733 (2022)

[134] Jan, Y.-T., Tsai, P.-S., Huang, W.-H., Chou, L.-Y., Huang, S.-C., Wang, J.-Z., Lu, P.-H., Lin, D.-C., Yen, C.-S., Teng, J.-P., *et al.*: Machine learning combined with radiomics and deep learning features extracted from ct images: a novel ai model to distinguish benign from malignant ovarian tumors. Insights into imaging **14**(1), 68 (2023)

[135] Cremonesi, F., Planat, V., Kalokyri, V., Kondylakis, H., Sanavia, T., Resinas, V.M.M., Singh, B., Uribe, S.: The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. Journal of biomedical informatics **141**, 104338 (2023)

[136] Groza, T., Gomez, F.L., Mashhadi, H.H., Muñoz-Fuentes, V., Gunes, O., Wilson, R., Cacheiro, P., Frost, A., Keskivali-Bond, P., Vardal, B., McCoy, A., Cheng, T.K., Santos, L., Wells, S., Smedley, D., Mallon, A.-M., Parkinson, H.: The international mouse phenotyping consortium: comprehensive knockout phenotyping underpinning the study of human disease. Nucleic Acids Research **51**(D1), 1038–1045 (2023) https://doi.org/10.1093/nar/gkac972

[137] AI for women's health group: AI in Women's Health: Bridging

research and patient voices (2025). https://www.crassh.cam.ac.uk/blog/ai-in-womens-health-bridging-research-and-patient-voices-i-event-report/