HHS Public Access

Author manuscript

Annu Rev Biomed Data Sci. Author manuscript; available in PMC 2024 August 10.

Published in final edited form as:

Annu Rev Biomed Data Sci. 2023 August 10; 6: 23–45. doi:10.1146/annurev-biodatasci-020722-105958.

Challenges and Opportunities for Data Science in Women's Health

Todd L. Edwards[1,2], **Catherine A. Greene**[2,3], **Jacqueline A. Piekos**[2,3], **Jacklyn N. Hellwege**[2,4], **Gabrielle Hampton**[1,2], **Elizabeth A. Jasper**[3,5], **Digna R. Velez Edwards**[2,3,6]

[1]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

[2]Vanderbilt Genetics Institute, Vanderbilt University Medicine Center, Nashville, TN

[3]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN

[4]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical center, Nashville, TN

[5]Center for Precision Medicine, Vanderbilt University Medicine Center, Nashville, TN

[6]Department of Biomedical Informatics, Vanderbilt University Medicine Center

Abstract

The intersection of women's health and data science is a domain that has historically trailed compared to other fields, but that has recently gained momentum. Much of this growth is being driven by new investigators who are moving into this area but also by the significant number of opportunities that have emerged in new methodologies, resources, and technologies in data science. Here, we describe some challenges, resources, and methods being used by data science investigators today and how these approaches are being used by women's health researchers to meet those challenges. We also describe the opportunities and limitations of applying these approaches for women's health outcomes and the future of the field, with emphasis on repurposing existing methodologies for women's health.

Keywords

data science; biomedical research; reproductive health; gynecologic health; electronic health records; women's health

Introduction

Women's health is the branch of medicine concerned with the study, diagnosis, and treatment of conditions that impact women, this includes both people of female biological sex and those who identify as female. Data Science is interdisciplinary and can be

Corresponding Author: Digna R Velez Edwards, 2525 West End Ave, Suite 600, 6th floor, rm 616, Nashville, TN 37203, digna.r.velez.edwards@vumc.org, Tel: 615.322.1288.

defined by several broad terms, here we discuss data science focused on women's health, encompassing big data, data mining, machine learning, and statistical approaches to gain knowledge of women's health from structured and unstructured data (1).

The pace of discoveries from scientific research that benefit women's health has been slowed by many factors including cultural biases, ethical dilemmas, misperceptions, inadequate resource allocation and development, and methodological challenges. However, opportunities are emerging to rapidly improve the quality and quantity of population-based women's health research using new approaches for amassing data, statistical analyses, and a growing workforce of researchers in the field (Figure 1).

Why Are There Differences By Sex?

Some health outcomes are directly influenced by biological sex. Sexual dimorphism in humans includes biological differences in many tissues and physiological processes, and health conditions often manifest with asymmetry. Humans also display sex-specific genetic architecture influencing many phenotypes, including susceptibility to certain diseases(2). Sexual differentiation begins early in development and has downstream effects on gene regulation and expression which is driven in part by differences in levels of circulating sex steroid hormones(3). These differences have been shown to influence hundreds of phenotypes(4) and are clinically relevant, as diseases may vary in susceptibility, pathogenesis, symptom severity, and response to treatment according to sex.

The origin of sex differences in disease has been hypothesized to be related to sex-specific evolutionary pressures exerted over the course of human evolution(5). Briefly, prehistoric patterns of human reproduction, characterized by high fecundity and high disease burden, necessitated maternal adaptation to a variety of immune threats that included external pathogens as well as the fetus itself. This increased immunomodulation compensated for the fetus's foreign genetic material while maintaining defenses against diseases before, during, and after repeated pregnancies. More recently, global industrialization has been associated with increases in breast and reproductive cancers as women's reproductive agency conflicts with ancestral adaptations to pregnancy(6). Such remnants of human evolution are also evident in sex-specific patterns of disease prevalence and pathophysiology.

What is Women's Health?

Women's health can be summarized by the following categories: 1. conditions of increased prevalence in women compared to men; 2. pregnancy-related health disorders; 3. gynecologic complications and cancers of specific tissues.

Conditions that Have Higher Prevalence in Women Compared to Men

Most common human diseases exhibit sex differences(2). Some of these disproportionately impact women despite occurring in both sexes. For instance, 80% of autoimmune diseases occur in women, affecting approximately 5% of women worldwide. These disorders, which include more than 70 conditions including multiple sclerosis, lupus, and rheumatoid arthritis, are mediated in part by the effects of sex steroid hormones on immune function(5;

7). The global prevalence of autoimmune diseases is rising in both sexes, though women's specific evolutionary immune adaptation to placental formation may account for the sex differences in prevalence for this class of disorders(5). Prevalence differences in disease are not necessarily static but may vary across the lifecourse. For example, the menopausal transition to lower estrogen and progesterone levels has been well defined as a change from decreased to increased risk of cardiovascular events compared to males(8).

Pregnancy-Related Health Disorders

Many conditions may be direct consequences of pregnancy and childbirth. We note that disorders of pregnancy may also impact pregnant people who are not female-identifying such as trans-men. Hypertensive disorders of pregnancy, for instance, complicate between 5% and 10% of pregnancies(9). Of these, preeclampsia has the highest morbidity and mortality, affecting between 5% to 7% of pregnant people worldwide each year while putting both mother and child at significant risk for further complications(10). Hypertensive disorders of pregnancy are associated with increased long-term risk of cardiovascular disease(9), contributing to sex disparities. These disorders have been associated with increased odds of stroke, Alzheimer's disease, and chronic kidney disease, among other conditions(11). The rapid and intense physiological changes associated with pregnancy and childbirth create unique health burdens that must be considered in discussions of women's health.

Gynecologic Complications and Cancers of Specific Tissues

Individuals with reproductive organs such as a uterus, vagina, fallopian tubes, and ovaries are at risk for health conditions impacting fertility and quality of life. Examples include endometriosis and uterine fibroids. Endometriosis is a chronic inflammatory condition where endometrial tissue grows outside the uterus which is estimated to affect at least 10% of reproductive-age women worldwide(12). Hallmarks include severely painful menstrual periods, depression, and potential infertility, with effects often exacerbated by delays in diagnosis and treatments(13). Uterine fibroids are benign tumors that can cause pain and prolonged periods, and recent evidence even implicates uterine fibroids as a possible cause of preterm birth(14). Prevalence varies by population but is higher in women of African descent. Overall, 7 in 10 women are expected to develop uterine fibroids during their reproductive years, posing a substantial health burden to women of reproductive age(15). Among other gynecologic disorders are polycystic ovarian syndrome (PCOS), pelvic organ prolapse, cervical dysplasia, and menstrual disorders.

Certain cancers are also predominant in or exclusive to women or those with female reproductive organs, such as breast cancer (sex-biased) and gynecologic cancers (sex exclusive; cervical, vaginal, ovarian, uterine [includes endometrial which begins in the uterus], vulvar, and fallopian tube cancer). In 2020, breast cancer made up almost a quarter of women's worldwide cancer diagnoses, with 2.3 million new diagnoses and 685,000 deaths(16). Health care interventions in developing nations have not yet matched efforts in affluent countries, which include breast cancer screening and adjusted recommendations such as increased physical activity or extended time breastfeeding to manage risk factors. Gynecologic cancers collectively made up almost a fifth of women's cancer diagnoses

globally in 2020(17). Cervical cancer is the most common of these and the fourth most frequent cancer among women overall, with an estimated 604,000 new cases and 342,000 deaths in 2020(18). It is also the only type of gynecologic cancer that can be prevented through screening, though women in developing countries often have limited access, resulting in higher mortality from cervical cancer. Prevention through vaccination against human papillomavirus (HPV) has proven both successful and cost-effective, encouraging further efforts for detection and elimination(18).

Women's health conditions are major global health concerns. Prevention of disease and improving health of women experiencing these conditions is a crucial goal for biomedical research. Opportunities exist to improve quality and quantity of knowledge in this domain by leveraging emerging resources and methods in data science.

Our Knowledge of Women's Health is Based on Data from Animal Models

As is the case with much of human disease biology, a great deal of progress in women's health is based on experimental animal models. An important and recent example is the development of the HPV vaccine, which utilized rabbits in preclinical studies(19). However, biomedical research has historically used male animals for non-pregnancy-related investigations, dismissing female test subjects due to concerns for how hormones might impact experimental results(4). Though changes to guidelines in recent years have emphasized inclusion of female animals and sex-disaggregation of data, decades of male-centric study design have had the consequence of sex differences being systemically understudied(20).

The mouse is the most frequently used model organism for studies of human pregnancy, though largely for convenience rather than suitability to women's health studies(21). Their large litters and short generation time makes them ideal for experimental science studies, but significant anatomical and molecular differences negatively impact generalizability to humans. Both mice and humans have hemochorial placentas, characterized by direct contact between maternal blood and the placental trophoblast layer. Mice also have a choriovitelline (yolk sac) placenta that persists until parturition. This structure is particularly problematic for studies of placental transfer of pharmaceutic or other chemical agents, as adverse effects or accumulation may be observed in mouse experiments that would not be relevant to humans(21).

Another important biological difference between humans and mice is that some protein hormones specific to the human placenta occur only within primates. Human chorionic gonadotropin (hCG), for instance, is essential for maintenance of the corpus luteum, as well as for immune-mediated maternal-fetal interactions within the placental bed(22). Human-murine genetic differences also include a human-specific gene cluster encoding prolactin/placental-lactogen-related genes, and a lack of other homologous regulatory microRNA genes (21).

As a result of these differences, it is difficult to reliably model pregnancy-related diseases in rodent models. Preeclampsia, for instance, does not occur naturally in mice(21).

Most experimental models, therefore, have preeclampsia induced by surgery or genetic manipulation, downstream of numerous initiating factors that cause the condition in humans(23). All available preeclampsia models have different benefits and shortcomings, as they are often developed to capture a specific aspect of preeclampsia yet fail to recapitulate the multisystemic disorder in its entirety. The RUPP rat, for example, is a popular model surgically induced to convincingly model hypertension, but fails to model some other diagnostic criteria for preeclampsia, such as proteinuria(23).

Non-human primate studies may be informative for some aspects of pregnancy and women's health, but are largely observational compared to experimental rodent models. Utilizing these models in research is helpful for investigating various candidate disease etiologies, but a more complete understanding relies on studies using human cells and tissues and on in-depth observational studies of human subjects. Although most mechanistic studies of pregnancy focus on murine models, there are some examples of primates such as the Japanese Macaque (Macaca fuscata) model which has been used to evaluate the impacts of high fat diets on offspring health (24; 25), and the Rhesus Macaque (*Macaca mulatta)*, where maternal infection during pregnancy has been studied (26). In summary, animal models are essential but can lack generalizability and should be utilized with an understanding of limitations to applicability to women's health.

With the rise of system biology approaches, animal model studies may find new utility for women's health research. Systems biology is computational and mathematical analysis of biological data collected from multiple biological systems. It can be used to mine and analyze multi-omics data (genomic, proteomic, and metabolomic analysis of biological samples including human and animal studies) across many studies and sources, as well as extract information regarding cells, tissues, and organism function obtained from individual experiments and data warehouses.

Data Science Study Design Considerations for Women's Health

Women's health research and studies that involve sex or gender research questions require additional considerations beyond the basics of study design. Design of human participant studies generally requires defining study age ranges and availability, prevalence, and measurement of exposures and outcomes, as well as considering ascertainment methods. Studies of women have unique concerns in nearly all of these areas: definitions of inclusion and exclusion criteria, recruitment, measurements, and analytic strategies (27). When defining comparison groups for study, researchers must consider female-specific life stages, such as menarche, pregnancy, and menopause, and the length of time spent in each life stage. Failure to account for these stages may result in incorrect conclusions.

Historical Context of Women's Health Research

Historically, research was conducted predominately or exclusively in men and merely generalized to women. Exclusion of women in studies was often justified based on concerns regarding exposures during childbearing years, avoidance of added complexities into study design and analysis by their inclusion, or the perspective that women were a vulnerable group and were at greater risk of being coerced into research(28).

In 1990, the U.S. General Accounting Office (GAO) published a report on the National Institute of Health (NIH) policies on the inclusion of women in study populations, drawing attention to the issue(29). As a result, the NIH created the Office of Research on Women's Health (ORWH) and announced the start of the Women's Health Initiative (WHI)(30–32). Several laws were also passed to increase women's inclusion and participation in research. The NIH Revitalization Act of 1993 (PL 103–43) instructed the NIH to establish guidelines for inclusion of women and minorities in clinical research and required the NIH to ensure that all clinical trials were conducted in a manner that allowed for analysis of effects in these subpopulations(33). The Food and Drug Administration also prioritized research that evaluated sex-based differences by establishing an Office of Women's Health in 1994(32).

Current Considerations for Data Science Studies

However, the legacy of excluding women in biomedical research is still evident years later. In 2001, the GAO published their findings on drugs removed from the market since 1997: eight of 10 prescription drugs removed from the U.S. market posed greater health risk for women than men(34). In 2014 in response to the historical exclusion of women in research the ORWH developed policy that required including a plan for how sex as a biological variable is addressed in the study design for all research grants submitted, including both human and animal studies(35). In 2016, the 21st Century Cures Act (PL 114–255) made further strides for women's health beyond previous requirements by requiring researchers conducting clinical trials to submit results of analyses stratified by sex, gender, race, and ethnicity to Clinicaltrials.gov(36). The Cures Act also directed NIH to consider whether those performing this research complied with this reporting requirement before awarding them additional grants in the future.

The requirements regarding sex as a biological variable pose several analytical considerations for the design of women's health research studies. Researchers must decide whether they are using sex and gender as an inclusion or exclusion criteria, a control variable, or in analyses aimed at specifically examining the effects of sex or gender on investigated outcomes. This consideration can greatly impact the required sample sizes and power of the study. Studies aiming to look at the impact of sex on outcomes must include enough women to be able to investigate differences in effects by sex.

Differences in life-stage and hormones across a woman's life present challenges in women's health research, though they can be accounted for in careful definition of the study population, study design, and use of appropriate statistical approaches. Time-varying or -dependent exposures or covariates, such as hormone levels, gestational weeks, or age, may require special time-varying covariance matrices in regression analyses. Statistical models, such as Cox regression, are also needed when analyzing time-to-event outcomes, like time to pregnancy in a study of fertility. Research on pregnancy has the added complexity of needing to account for pre-pregnancy factors, gestational age, the physiological differences that occur during specific trimesters, and/or fetus-specific effects (e.g., genetic variation) in addition to accounting for characteristics of the mother.

Although inclusion and recruitment of women has increased in recent years, women as study participants are still underrepresented in many biomedical research fields(27). Significant

work is needed to improve health equity and account for differences in disease risk, presentation, pathophysiology, and treatment response based on sex. Factors influencing women's participation in clinical research include how well the research is explained, risk of unknown side effects, language barriers, familial responsibilities, comorbidities, and inconveniences(37–39). Risk to the fetus is highly cited as a barrier to participation for pregnant persons(38). Clear communication in participant-researcher interactions and study materials, including consent documents, is necessary to increase women's participation in research. Community engagement approaches can aid in building trusting relationships, an often-cited facilitator to recruitment and retention(40). Use of social media should also be considered when designing study recruitment strategies. Flexibility in scheduling, frequent communication, and culturally sensitive practices can aid in retention of both pregnant and non-pregnant people(40).

The WHI's studies provide an example of how inclusion and exclusion criteria, such as life stage and length of time in the stage, can lead to different conclusions. One arm of WHI, which enrolled postmenopausal women, investigated the risk and benefits of hormone replacement therapy (HRT). After halting the study before completion, WHI found estrogen therapy resulted in an increased risk of heart disease and breast cancer(41). However, largely due to their interest in cardiovascular and cancer outcomes which are more common in older individuals, the study largely enrolled women who were older than those who would normally be considered for initiating HRT in real world clinical practice. Women were often more than 10 years from the start of menopause.

Reanalysis of the WHI trial data, as well new studies, have since demonstrated the benefits of HRT for younger women (50–59 years) or those in in the early postmenopausal (within 10 years of menopause onset)—finding reduced coronary diseases and all-cause mortality in women from these stages on HRT contrasted to a comparable group of women not on HRT(42–47). A large controlled trial from Denmark also demonstrated reduced risks of heart disease and death from heart disease in healthy women who took combined HRT for 10 years immediately after menopause(48). Thus, careful consideration of comparison groups and real-world applications are necessary when defining comparison groups in women's health research as incorrect or poorly defined study populations can have profound consequences in terms of study findings and significance.

The Intersectional Nature of Women's Health Conditions Create Challenges for Data Science

Women's health research often intersects with health conditions that disproportionately impact underrepresented populations such as racial and ethnic minorities, creating additional challenges. The clearest examples are studies of pregnancy health where Black women are at the highest risk of dying in childbirth, experiencing a pregnancy loss, or experiencing adverse pregnancy complications such as a preterm delivery compared to all other racial groups (49–51). The disparities are heightened by structural factors such as stereotyping (50), insurance access, and accessible healthcare, including too few community healthcare clinics(52). There are efforts to reduce disparities through improved access to contraception, increased prenatal care, and increased STD and HIV screening and treatment(53).

Other examples of intersectional categories include sex and gender minorities, such as transgender women who face unique mental health and healthcare challenges that are difficult to capture completely using only resources such as electronic health records (EHR). From a data science perspective, these intersectional categories of women's health create challenges in terms of the power of studies focused on intersectional risk populations and designing studies that capture the complex network of factors contributing to outcomes. For example, studies conducted with EHR will lack information on social determinants of health: socioeconomic factors, systemic racism, some mental health stressors, nutrition, and behaviors such as smoking, drug, and alcohol use. EHRs also are poorly able to capture risk contributions due to the health care systems itself, such as those due to access to care (14).

Data Science Studies in Women's Health Are Primarily Performed in Population-Based Data

Existing studies of women's health using data science approaches come from a wide variety of study design frameworks: population-based cohorts that include state and country-wide registries, EHR repositories (some of which have affiliated biobanks), observational cohort studies, case-control studies, and secondary analyses of trial data; each study type has strengths and weaknesses. The design selected depends on the study purpose or question and must consider the availability and timing of assessment for exposures and outcomes, the study population, time, and cost. We briefly highlight the major types of studies with designs that allow for testing of hypotheses with particular emphasis of those generating large volumes of data most suitable for data science approaches.

Electronic Health Record (EHR) Repositories

Large quantities of data related to health, clinical practice, treatment efficacy and safety, medication use, demographics, vital measures, and other domains relevant to women's health research are routinely collected as part of the clinical enterprise. There are also a growing number of biorepositories and EHR-linked biobanks that are being used for studies of women's health outcomes (54). Use of extant EHR data for research can provide advantages over other study designs such as reduced staffing, costs, and shorter time to completion compared with recruiting cohorts(55), as well as the benefits of no recall bias (since data are collected in real time), large sample sizes, population representation, and detailed prescribing information (56). However, use of these data also requires overcoming several design and methodological challenges, in part because EHR systems are typically not designed to support research(57). Some disadvantages can include lack of detailed information about lifestyle, geography, economic factors, familial morbidities, health history, or other relevant information for many research questions.

Several recent reviews and perspective pieces in the literature have described advantages and challenges of working with EHR data. Considerations for EHR-based studies have been recently reviewed (55–57) (55). Key steps for using EHR data for population-based research are: cohort building, defining variables, feature selection, study design, and results validation. Each of these areas has distinct challenges as well. Awareness of sample selection bias in cohort building and definition of appropriate patient subpopulations,

including identification of primary care providers are critical considerations to make when initiating a new study. Variable selection and definition also may be complicated depending on if they are derived from structured or unstructured information and may also suffer from imprecise variable definitions and limitations to algorithms. Additional areas of concern are confounding between density of EHR activity and disease severity, accounting for temporal changes in health system composition and treatments, differential provider tendencies for treatments, and distinguishing prevalent from incident cases. Validation of results may be particularly hampered by each one of these challenges, as well as structural differences between EHR platforms used across medical centers/clinical enterprises.

Each study in the EHR requires a unique set of considerations to deliver reliable inferences. Guidelines such as STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) and RECORD (REporting of studies Conducted using Observational Routinely-collected health Data) exist to provide best practices for reporting results from observational studies using health data(58–60). Consultation with experienced investigators can often provide many insights and avoid or mitigate the effects of some of the problems listed above.

In the context of women's health research, EHR data can be used to study conditions that might be difficult to measure in cohort studies. An example of this is uterine fibroids, where as many as 51% of women are misclassified without confirmation by imaging(61). Providing ultrasounds to thousands of participants in a cohort study would be expensive and time-consuming, whereas those data exist in many health systems as imaging reports, procedure codes, and diagnoses. We have developed and validated a phenotyping algorithm for the identification both of uterine fibroid cases and confirmed fibroid-free controls(62). Our algorithm requires visual confirmation of UF status through pelvic imaging (cases and controls) or surgery (cases).

State and Country-Wide Registries

Clinical data collected from state and country-wide registries are utilized for studies of women's health, often to assess changes and trends in women's health and pregnancy outcomes at the state and country level. Examples of types of state registries include vital records data, state health department registries, and cancer/tumor registries. Examples of country-wide cohorts are the UK BioBank(63), the Estonian Biobank cohort(64), Biobank Japan(65), the Danish Medical Birth Registry(66), and the FinnGen study(67), several of which also have integrated EHR data available through their comprehensive national healthcare systems.

Observational Study Designs

Cross-Sectional Study Design: Cross-sectional study designs assess both exposure status and presence or absence of disease at the same point in time(68). Under this design, investigators are unable to determine whether a temporal relationship exists between exposures and outcomes. Thus, they are often considered the weakest observational study design because of their limited capacity for establishing causality.

Case-Control Study Design: Case-control designs begin after both exposures and disease have occurred but only the disease status of individuals is known at the beginning of the study. They allow for comparison of the proportions of individuals who have experienced the exposure of interest in individuals with the phenotype of interest (cases) to those without the phenotype of interest (controls). This study design is ideal for less common outcomes and can often be completed faster and cost less than cohort study designs.

Cohort Study Design: Cohort studies begin with groups of individuals whose exposure status is known, though some may start with a defined population which is selected based on a factor not related to exposure with some individuals becoming exposed and others remaining unexposed. Disease status is unknown at the time of cohort creation. The incidence of disease can be calculated in both the exposed and nonexposed groups, allowing for comparison through calculation of relative risk. Cohort designs are desirable when the exposure of interest is rare. There are two major types of cohort designs, briefly discussed below.

Prospective cohorts begin with establishment of a cohort for which exposure status is known. The cohort is then followed into the future to determine their disease status. This longitudinal design is powerful, as it minimizes potential for biases in assessing the exposure and, generally, allows for more accurate assessments of temporal relationships between exposures and outcomes. However, they are often expensive and time-consuming.

A retrospective cohort is devised after both exposures and some or all cases of diseases have occurred, though investigators do not know individuals' disease status prior to beginning the study. They require historical data to identify and assemble a cohort with the relevant exposed and unexposed groups. The cohort is followed over a set period either in the past or until present-day to determine risk of disease in both exposed and unexposed groups. Though they can be completed more efficiently than prospective studies and are usually less expensive, retrospective cohorts rely on the availability and consistency of past exposure data.

Other Observational Study Designs: Mixed longitudinal cohorts involve both prospective and retrospective follow-up components. Nested case-control and case-cohort designs combine elements of cohort and case-control studies, though selection of the control group differs between the design. Case-crossover designs are useful in the study of acute time-varying exposures that produce transient changes in risk of a phenotype within a short period of time. They compare the case's exposure status immediately before its occurrence with their exposure status at a prior time.

Intervention (Experimental) Studies

Intervention studies typically provide the best quality of evidence for causal relationships between exposures and phenotypes. They include explanatory and pragmatic trials. Though intervention studies have huge advantages, they often lack generalizability, require subjects comply with study procedures, and have large time, sample size, and financial costs.

Clinical Trials: Clinical trials are arguably the most recognized form of intervention studies. They are planned experiments designed to assess the safety and efficacy of an intervention by comparing outcomes in a group of individuals treated with a test intervention with those observed in a comparable group of individuals receiving a control intervention, where both groups are enrolled, receive an intervention, and are followed over the same period (69).

Pragmatic trials: Pragmatic trials are used to evaluate the efficacy and effectiveness of interventions in real-world scenarios. They can be beneficial when interventions or treatments are already available, but researchers wish to test its overall effectiveness in routine practice. Thus, while they do not allow investigators to study contributions of different components of care, they provide evidence that interventions work in real-life, not just in tightly controlled clinical trial settings.

Other Intervention Studies: Special cases of clinical trials can also be used to study women's health. Cross-over designs are special cases of randomized trials where each subject serves as their own control, receiving both the intervention and control during the study. This design is useful for outcomes that are transient and has the advantages of reducing both sample size needed for the study and variability. Factorial designs also employ randomization but allow for comparison of a combination of interventions and have the potential to shorten the time it takes to conduct a trial.

The Recent Growth of Published Data Science Studies on Women's Health

We conducted a PubMed search of the literature to assess growth of publications in the areas of data science and women's health from 1967 (earliest available search) to 2022, as well as patterns of growth in subcategories of data science, including research within clinical EHR and genomic studies. Below is a summary of the search and the findings.

We used the following medical subject headings (MeSH) terms and filtering criteria to conduct our search of the literature in PubMed. Our MeSH terms were the following: (((((“Women”[MeSH]) OR (“Women’s Health”[MeSH])) OR (“Women’s Health Services”[MeSH])) OR (woman)) OR (women)) AND ((((((((((“Data Science”[MeSH]) OR (“Computational Biology”[MeSH])) OR (“Medical Informatics”[MeSH])) OR (“Informatics”[MeSH])) OR (“Biostatistics”[MeSH])) OR (“Statistics as Topic”[MeSH])) OR (“Algorithms”[MeSH])) OR (“Genomics”[MeSH])) OR (“Molecular Epidemiology”[MeSH])) OR (computational genetics)) OR (genetics epidemiology)) OR (computational epidemiology)). We also filtered the search by year, study type, and limited to human studies. The types of studies included in the search included Clinical Study, Clinical Trial, Clinical Trial, Phase I, Clinical Trial, Phase II, Clinical Trial, Phase III, Clinical Trial, Phase IV, Comparative Study, Controlled Clinical Trial, Meta-Analysis, Multicenter Study, Observational Study, Pragmatic Clinical Trial, Randomized Controlled Trial, Twin Study, Validation Study, Humans, from 1967/1/1 – 2022/10/03.

We identified 101,623 total human studies across our time range that included data science and women's health topics after applying our filtering criteria (Figure 2). There has been a steady increase in data science publications in women's health since 2000, with the peak occurring in 2019 (Figure 2). The low number for 2022 is because our search collected data through October 2022 and the year is incomplete as of this writing, as well as some publications not necessarily becoming immediately indexed in PubMed upon publication. It is unknown why there was a slight drop in the number of published studies between 2020 and 2021. We speculate this may be due to the COVID-19 pandemic and the resultant drop in productivity by researchers due to temporary shutdowns and transitions into virtual and hybrid work models. It could also be due to the disproportionate impact of pandemic on the female workforce. Studies have shown(70) that there is a relationship between women's health and women's leadership in academic medicine.

When we further subdivided the publications by studies mentioning use of electronic health record (EHR) data or genomics we identified 1,007 and 8,012 studies, respectively (Figure 3A and 3B). There was a consistent rise in published women's health-focused EHR and genomic studies since the early 2000's peaking in 2021, which is not surprising considering the expanded growth and use of EHRs in health care systems and both the completion of the *Human Genome Project* (completed in 2003) and emergence of large-scale genomics technology (first genome-wide association study published in 2005) over the same period.

Machine Learning and Artificial Intelligence Approaches

Artificial intelligence (AI) is a branch of computer science in which intelligence that is demonstrated by humans or other animals is simulated by computers. Examples of this are tasks such as visual perception, automated problem solving, and natural language processing (NLP). NLP phenotyping methods have been developed for research(71) and have been applied to many problems, such as identifying differences within cancer types, with examples in ovarian and breast tumors(72).

Machine Learning (ML) is a discipline within AI which develops software that can learn autonomously. Expert systems and data mining are common applications of ML technology, and methods such as neural networks and genetic algorithms are common strategies for implementing ML.

Applications of AI in health research include rapidly processing EHRs to identify evidence of a disease from clinical notes or evaluating many images for evidence of a particular visual feature. ML and AI approaches have been used to classify participants for women's health phenotypes in EHR studies. Examples include studies of spontaneous preterm birth (73), gestational diabetes and pre-eclampsia (74), and pregnancy complications(75; 76). Davidson et al. provides a comprehensive review of ML and AI methods applied to various stages and conditions related to pregnancy(77). These studies illustrate a recent trend toward using more sophisticated automated approaches for cohort construction, compared with earlier methods that relied on human expert knowledge and intuition for algorithm development.

Omics in Women's Health

"-Omics" is a suffix indicating the study of large amounts of biological data. There are many types of -omics: genomics, metagenomics, microbiomics, epigenomics, transcriptomics, proteomics, and metabolomics.

Next-generation sequencing and genotyping technologies have advanced the field of genomics by facilitating cheaper and faster sequencing and genotyping of large cohorts. The large amount of genetic data is then used in genome-wide association studies (GWAS) where hundreds of thousands up to hundreds of millions of genetic variants are interrogated for association with an outcome(78). Breast cancer is an example of how genomic technologies have led to discovery of over 150 associated loci(79). Researchers have leveraged this genomic information to understand more about the etiology of disease through functional studies, build genetic scores to identify high risk individuals, and enhance precision medicine by matching therapeutics to patients based on their genetic profile(80). The results of most GWAS are compiled in the National Human Genome Research Institute-European Bioinformatics Institute GWAS Catalog(81).

The epigenome is made up of modifications to chromatin structure which affect DNA transcription and alter gene expression. The epigenome can be interrogated through genome-wide DNA methylation and histone modification analyses. Epigenome studies in cancer and cancer-like phenotypes such as uterine fibroids have been productive in understanding tumor biology through observing processes that result in aberrant expression of genes and identification of possible druggable epigenetic marks(82).

The study of globally transcribed (expressed) genes within a cell, tissue, or individual is the transcriptome. A commonly used technique is RNA sequencing which quantifies the number of transcripts for each gene in a tissue sample. This technique provides information that can be leveraged in numerous ways, one of which is to detect differentially expressed genes (DEG) between diseased and normal tissues. DEGs can indicate the molecular pathways that are dysregulated in disease states. A review of RNAseq experiments on preeclampsia identified 250 DEGs between placentas from preeclamptic and healthy pregnancies(83). Incorporating these newly identified genes with systems biology approaches, researchers were able to build protein-protein interaction networks and identified extracellular matrix organization and immune processes as biological processes dysregulated by preeclampsia (83).

There are multiple publicly available resources for transcriptomic data from both tumor and normal tissue. The Cancer Genome Atlas (TCGA) is a landmark program from the US National Cancer Institute that contains data on matched cancer and normal samples for 33 tumors from over 11,000 patients collected over a 12-year period. These data are public and accessible via the TCGA website. The International Cancer Genome Consortium data portal is another public resource that is designed to provide visualizations, analysis, and interpretation of large catalog of mutations in a diverse set of tumors(84). Another large-scale transcriptomics program is the Genotype-Tissue Expression (GTEx) Project, supported by the NIH. The goal of GTEx was to create a map of genetic determinants

of gene expression (expression quantitative trait loci [eQTLs]). The most recent version of this resource contains eQTL information from 54 human tissues of 948 mostly postmortem donors(85; 86). These data can be leveraged in concert with GWAS to conduct transcriptome-wide association studies (TWAS) in which associations between genetically-predicted gene expression levels and outcomes can be statistically inferred(87–90). TWAS have been conducted for uterine fibroids(91; 92), breast cancer (93–95), age at menopause (96), mammographic density(93), postpartum depression(97), age at menarche(98), ovarian cancer(99; 100), and other outcomes.

Proteomics is the study of protein expression in a given disease, tissue, or individual. Analysis of proteins as biomarkers has the advantage of having the final three-dimensional structure and any post-translational modifications, rather than the relying on inferences from transcript precursors used in transcriptomics. However, proteins are more spatially localized than their transcriptome predecessors and therefore may be more difficult to assay comprehensively. Proteomics has been used as diagnostic biomarkers for gestational diabetes mellitus(101), in association studies of ovarian cancer(102; 103), follicular fluid(104), and spontaneous abortion(105), among others.

Metabolomics is the global profile of metabolites from a given cell, tissue, or organism. Measured metabolites often may arise from both exogenous compounds as well as those produced internally. Metabolomic studies have identified abnormal metabolism of lipids, amino acids, carbohydrates, and steroid hormones in polycystic ovarian syndrome in search of diagnostic markers and drug targets(106), measured endometrial receptivity in recurrent miscarriage(107), endometrial cancer screening(108), pelvic organ prolapse(109), and many other outcomes.

Risk Prediction Modeling

An essential aspect of implementing precision medicine in women’s health is accurately assessing risk. This is accomplished by developing a model for calculating the posterior probability of an outcome by conditioning on observed values of several variables. Statistical and computer science approaches exist for model construction, which generally consists of evaluating a set of candidate data elements or features, selecting a subset of features, and assigning quantifications of importance (weights) to each feature in the model. Many data types can be included as features in evaluation of risk prediction, not limited to core demographic information, environmental or lifestyle exposures, biomarkers, health histories, and genomic and other -omic measures.

Study designs for developing predictive models must incorporate strategies to mitigate overfitting, which is a modeling error where the model describes an outcome in a specific sample very well but does not generalize to other samples. This can occur due to random type I errors that arise due to sampling stochasticity and selecting too many features as predictors with weights that correspond more to sampling error than reality. Both statistical and machine learning strategies for model construction can result in overfitting. Common approaches used to reduce overfitting are using mutually exclusive training and testing data, and resampling strategies like cross-validation and bootstrapping to estimate the

performance of a model in independent data. Additionally, penalizing the likelihood function in regression models to encourage model parsimony can mitigate overfitting as well.

Predictive models that perform well can be implemented as clinical decision support tools that alert clinicians when a patient's risk profile indicates they may have or are likely to develop a condition. This can improve operational efficiency by reducing costs and potentially providing better outcomes for patients.

Validated predictive models have been developed using various approaches for many women's health conditions. There are many reviews describing the numerous models available for such conditions as caesarian section complications(110), successful vaginal birth after caesarian section(111; 112), AI-based predictions of breast cancer recurrence(113), ML approaches in predicting postpartum depression (114), postpartum hemorrhage(115), natural menopause onset(116), among others. Many of these studies rely in part or entirely on clinical data from EHRs.

Polygenic Risk Score Modeling

Polygenic scores are composite genetic variables that summarize the heritable component of risk of a given outcome. These scores are usually calculated using the results from GWAS, where a linear combination of estimated effects of influential loci from each region of the genome in combination with observed genotypes are used to calculate a score for each participant in a different study. To perform this calculation, both dense genotyping for study subjects and previous GWAS on independent participants are required to address overfitting. Many methods exist to perform the selection and weighting of genetic features from GWAS results, and this is an area of brisk methodological development.

This approach has been applied to several diseases including PCOS(117), endometriosis(118), uterine fibroids(119), and epithelial ovarian cancer(120). Multiple polygenic scores have been developed for breast cancer, a review of which can be found in (121). The multifactorial Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) model for predicting future breast cancer risks(122; 123) is an approach that has been incorporated in several clinical management guidelines in the UK, North America, and other countries. BOADICEA incorporates a validated polygenic score with clinical factors, pathogenic variants, family history, a lifestyle, hormonal, and reproductive risk factor survey, and demographic factors. In a recent evaluation, the polygenic score was found to contribute most to risk stratification of all predictors(124).

Collaborative and Network-Based Research Studies

Researchers in large-scale biomedical science have developed consortium models for collaboration among independent investigators to amass resources and provide opportunities to conduct multi-site, multi-stage study designs. This is particularly important in women's health areas as we strive to overcome historical underrepresentation. An example of this is the pelvic floor disorders network, a research program established by the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Another example is

the World Endometriosis Research Foundation's Endometriosis Phenome and Biobanking Harmonization Project led by the University of Oxford, where academic and biotechnology industry scientists develop standardized data collection instruments and sample collection protocols in endometriosis research. Many other women's health research networks exist in other domains. Support for these collaborations is highly varied and is sometimes based on programmatic goals from private foundations, government agencies, or academic institutions, while others consist of individually funded researchers contributing resources from their ongoing studies.

A domain where the network approach is developing rapidly is in genetic epidemiology. Multi-site, multi-stage study designs have been applied to women's heath traits in uterine fibroids(91; 92; 125), PCOS (126), endometriosis(127), pelvic organ prolapse (128), preeclampsia (129), pre-term birth (130), and other conditions. The reason for the rapid growth of these collaborative networks is both pragmatic and based on underlying biology. For most common diseases, the genetic architecture involves many locations in the genome with relatively subtle effects. Additionally, evaluating the entire genome for association between common variation and outcomes requires many statistical tests, which necessitates strict multiple testing criteria that reduce statistical efficiency. The result of all these factors is that very large sample sizes are often required to elucidate the genetic architecture of common traits. As investigators realized the limitations of conducting investigations in single studies, collaborations developed among them to increase statistical power, often by combining evidence for association at each genetic location from multiple studies using meta-analysis.

Building a Larger Workforce of Women's Health Data Science Researchers

The disparities in funding women's health focused research create challenges in growing the workforce in this area. Studies evaluating funding patterns by the NIH have shown that there are disparities in funding disease that disproportionately impact women, with funding patterns favoring diseases that are specific to males (131). Furthermore, studies demonstrate that studies of male-focused diseases receive twice the funding compared with diseases more prevalent or specific to females (132). The issue is compounded by funding differences by sex of the principal investigators. A study conducted by the Northwestern Institute on Complex Systems Army Research Lab identified that out of 53,000 grants awarded between 2006 and 2017 new women principal investigators receiving their first funding awards were awarded 24 percent less than new male principal investigators (133). There is also disparity in the number of women principal investigators applying for NIH-funded grants. According to the deputy director of the National Institute of General Medical Science, less than one-third of first-time applicants for NIH grants are women. For example, in 2015, women received 53% of biology Ph.D.'s, held 44% of assistant professorships, and only 35% of the professoriate with PhDs in biology (134). The study also identified pay disparities and challenges for women matriculating and finishing graduate school and post-doctoral fellowships(134; 135).

Challenges persist for women researchers and research in women's health. There are active endeavors aimed at reducing burdens such funding opportunities, access to healthcare,

access to affordable childcare, and overall attitude toward women investigators in the scientific community, and success of these measures will promote a more inclusive environment and benefit all women through successes in women's health research.

Conclusions

Population-based science using data science approaches is a rapidly growing and developing aspect of women's health research. In our view, it is an essential element of developing precision medicine to reduce the burden of disease in women. Research in women's health has previously been slowed by ethical, cultural, and logistical obstacles. However, the development of large-scale electronic health record databases and biobanks can alleviate many of those challenges. The issue of processing these resources to derive the benefits for the field can be addressed by developing the workforce and building the community of data scientists in women's health. Expanded career development awards from the National Institutes of Health and other enterprises for early career scientists to pursue these goals will be necessary to take advantage of the opportunities for discovery, innovation, and translational impact that are available in women's health.

References

1. Cao L 2017. Data Science: A Comprehensive Overview. ACM Computing Surveys 50:43:1
2. Ober C, Loisel DA, Gilad Y. 2008. Sex-specific genetic architecture of human disease. Nat Rev Genet 9:911–22 [PubMed: 19002143]
3. Arnold AP. 2017. A general theory of sexual differentiation. J Neurosci Res 95:291–300 [PubMed: 27870435]
4. Karp NA, Mason J, Beaudet AL, Benjamini Y, Bower L, et al. 2017. Prevalence of sexual dimorphism in mammalian phenotypic traits. Nat Commun 8:15475 [PubMed: 28650954]
5. Natri H, Garcia AR, Buetow KH, Trumble BC, Wilson MA. 2019. The Pregnancy Pickle: Evolved Immune Compensation Due to Pregnancy Underlies Sex Differences in Human Diseases. Trends Genet 35:478–88 [PubMed: 31200807]
6. Fathalla MF. 2019. Impact of reproductive evolutionary mismatch on women's health and the need for action and research. Int J Gynaecol Obstet 144:129–34 [PubMed: 30341890]
7. Whitacre CC. 2001. Sex differences in autoimmune disease. Nat Immunol 2:777–80 [PubMed: 11526384]
8. Willemars MMA, Nabben M, Verdonschot JAJ, Hoes MF. 2022. Evaluation of the Interaction of Sex Hormones and Cardiovascular Function and Health. Curr Heart Fail Rep 19:200–12 [PubMed: 35624387]
9. Ying W, Catov JM, Ouyang P. 2018. Hypertensive Disorders of Pregnancy and Future Maternal Cardiovascular Risk. J Am Heart Assoc 7:e009382 [PubMed: 30371154]
10. Rana S, Lemoine E, Granger JP, Karumanchi SA. 2019. Preeclampsia: Pathophysiology, Challenges, and Perspectives. Circ Res 124:1094–112 [PubMed: 30920918]
11. Mauvais-Jarvis F, Bairey Merz N, Barnes PJ, Brinton RD, Carrero JJ, et al. 2020. Sex and gender: modifiers of health, disease, and medicine. Lancet 396:565–82 [PubMed: 32828189]
12. 2021. Endometriosis https://www.who.int/news-room/fact-sheets/detail/endometriosis#:~:text=Endometriosis%20affects%20roughly%2010%25%20(190,and%20girls%20globally%20(2
13. Chapron C, Marcellin L, Borghese B, Santulli P. 2019. Rethinking mechanisms, diagnosis and management of endometriosis. Nat Rev Endocrinol 15:666–82 [PubMed: 31488888]
14. Landman A, Don EE, Vissers G, Ket HCJ, Oudijk MA, et al. 2022. The risk of preterm birth in women with uterine fibroids: A systematic review and meta-analysis. PLoS One 17:e0269478 [PubMed: 35653408]

15. Stewart EA, Cookson CL, Gandolfo RA, Schulze-Rath R. 2017. Epidemiology of uterine fibroids: a systematic review. BJOG 124:1501–12 [PubMed: 28296146]
16. 2021. Breast Cancer https://www.who.int/news-room/fact-sheets/detail/breast-cancer
17. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, et al. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–49 [PubMed: 33538338]
18. 2022. Cervical cancer https://www.who.int/news-room/fact-sheets/detail/cervical-cancer
19. Mejia AF, Culp TD, Cladel NM, Balogh KK, Budgeon LR, et al. 2006. Preclinical model to test human papillomavirus virus (HPV) capsid vaccines in vivo using infectious HPV/cottontail rabbit papillomavirus chimeric papillomavirus particles. J Virol 80:12393–7 [PubMed: 17005666]
20. Danska JS. 2014. Sex matters for mechanism. Sci Transl Med 6:258fs40
21. Schmidt A, Morales-Prieto DM, Pastuschek J, Frohlich K, Markert UR. 2015. Only humans have human placentas: molecular differences between mice and humans. J Reprod Immunol 108:65–71 [PubMed: 25817465]
22. Carter AM. 2022. Evolution of Placental Hormones: Implications for Animal Models. Front Endocrinol (Lausanne) 13:891927 [PubMed: 35692413]
23. Gatford KL, Andraweera PH, Roberts CT, Care AS. 2020. Animal Models of Preeclampsia: Causes, Consequences, and Interventions. Hypertension 75:1363–81 [PubMed: 32248704]
24. Elsakr JM, Zhao SK, Ricciardi V, Dean TA, Takahashi DL, et al. 2021. Western-style diet consumption impairs maternal insulin sensitivity and glucose metabolism during pregnancy in a Japanese macaque model. Sci Rep 11:12977 [PubMed: 34155315]
25. Elsakr JM, Dunn JC, Tennant K, Zhao SK, Kroeten K, et al. 2019. Maternal Western-style diet affects offspring islet composition and function in a non-human primate model of maternal over-nutrition. Mol Metab 25:73–82 [PubMed: 31036449]
26. Boktor JC, Adame MD, Rose DR, Schumann CM, Murray KD, et al. 2022. Global metabolic profiles in a non-human primate model of maternal immune activation: implications for neurodevelopmental disorders. Mol Psychiatry
27. Institute of Medicine (US) Committee on Women's Health Research. 2010. Methodologic issues in women's health research. In Women's Health Research: Progress, Pitfalls, and Promise Washington (DC): National Academies Press (US). Number of.
28. Mazure CM, Jones DP. 2015. Twenty years and still counting: including women as participants and studying sex and gender in biomedical research. Bmc Womens Health 15
29. Nadel MV. 1990. National Institutes of Health: problems in implementing policy on women in study populations. In Subcommittee on Housing and Consumer Interest, Select Committee on Aging Washington, DC: United States General Accounting Office
30. The Women's Health Inititative Study Group. 1998. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials 19:61–109 [PubMed: 9492970]
31. Pinn VW. 1994. The role of the NIH's Office of Research on Women's Health. Acad Med 69:698–702 [PubMed: 8074758]
32. https://www.fda.gov/about-fda/office-commissioner/office-womens-health
33. National Institutes of H 1993. NIH revitalization act of 1993 (PL 103–43). Subtitle B. Sections 131:133
34. Heinrich J 2001. Drug safety: most drugs withdrawn in recent years had greater health risks for women. ed. THOJS A letter to The Honorable Tom Harkin, The Honorable Barbara A. Mikulski, United States Senate, The Honorable Henry Waxman, House of Representatives Washington, DC: United States General Accounting Office,
35. Clayton JA, Collins FS. 2014. Policy: NIH to balance sex in cell and animal studies. Nature 509:282–3 [PubMed: 24834516]
36. 2016. Public Law 114 – 255 - 21st Century Cures Act. ed. Office of the Federal Register National Archives and Records Administration: U.S. Government Publishing Office

37. Courvoisier N, Storari C, Lesage S, Vittoz L, Barbieux C, et al. 2022. Facilitators and barriers of women's participation in HIV clinical research in Switzerland: A qualitative study. HIV Med 23:441–7 [PubMed: 35178844]
38. Myles S, Tocci C, Falk M, Lynch S, Torres C, et al. 2018. A Multicenter Investigation of Factors Influencing Women's Participation in Clinical Trials. J Womens Health (Larchmt) 27:258–70 [PubMed: 29148879]
39. van der Zande ISE, van der Graaf R, Hooft L, van Delden JJM. 2018. Facilitators and barriers to pregnant women's participation in research: A systematic review. Women Birth 31:350–61 [PubMed: 29373261]
40. Goldstein E, Bakhireva LN, Nervik K, Hagen S, Turnquist A, et al. 2021. Recruitment and retention of pregnant women in prospective birth cohort studies: A scoping review and content analysis of the literature. Neurotoxicol Teratol 85:106974 [PubMed: 33766723]
41. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, et al. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 288:321–33 [PubMed: 12117397]
42. Boardman HM, Hartley L, Eisinga A, Main C, Roque i Figuls M, et al. 2015. Hormone therapy for preventing cardiovascular disease in post-menopausal women. Cochrane Database Syst Rev:CD002229
43. Cagnacci A, Venier M. 2019. The Controversial History of Hormone Replacement Therapy. Medicina (Kaunas) 55
44. Manson JE, Chlebowski RT, Stefanick ML, Aragaki AK, Rossouw JE, et al. 2013. Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. JAMA 310:1353–68 [PubMed: 24084921]
45. Rossouw JE, Prentice RL, Manson JE, Wu L, Barad D, et al. 2007. Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. JAMA 297:1465–77 [PubMed: 17405972]
46. Salpeter SR, Walsh JM, Greyber E, Ormiston TM, Salpeter EE. 2004. Mortality associated with hormone replacement therapy in younger and older women: a meta-analysis. J Gen Intern Med 19:791–804 [PubMed: 15209595]
47. Salpeter SR, Walsh JM, Greyber E, Salpeter EE. 2006. Brief report: Coronary heart disease events associated with hormone therapy in younger and older women. A meta-analysis. J Gen Intern Med 21:363–6 [PubMed: 16686814]
48. Schierbeck LL, Rejnmark L, Tofteng CL, Stilgren L, Eiken P, et al. 2012. Effect of hormone replacement therapy on cardiovascular events in recently postmenopausal women: randomised trial. BMJ 345:e6409 [PubMed: 23048011]
49. Salsberry PJ, Reagan PB, Fang MZ. 2013. Disparities in Women's health across a generation: a mother–daughter comparison. Journal of Women's Health 22:617–24
50. Sutton MY, Anachebe NF, Lee R, Skanes H. 2021. Racial and ethnic disparities in reproductive health services and outcomes, 2020. Obstetrics and Gynecology 137:225 [PubMed: 33416284]
51. Hornbuckle LM, Amutah-Onukagha N, Bryan A, Skidmore Edwards E, Madzima T, et al. 2017. Health disparities in women p. 1179562X17709546: SAGE Publications Sage UK: London, England
52. Ranji U, Rosenzweig C, Salganicoff A. 2018. Women's Coverage, Access, and Affordability: Key Findings from the 2017 Kaiser Women's Health Survey. Issue Brief, Kaiser Family Foundation, available at: https://www.kff.org/womens-health-policy/issue-brief/womens-coverage-access-and-affordability-key-findings-from-the-2017-kaiser-womens-health-survey
53. Johnston EM, Strahan AE, Joski P, Dunlop AL, Adams EK. 2018. Impacts of the Affordable Care Act's Medicaid expansion on women of reproductive age: differences by parental status and state policies. Women's Health Issues 28:122–9 [PubMed: 29275063]
54. Hallam L, Vassallo A, Pinho-Gomes A-C, Carcel C, Woodward M. 2022. Does Journal Content in the Field of Women's Health Represent Women's Burden of Disease? A Review of Publications in 2010 and 2020. Journal of Women's Health 31:611–9
55. Sauer CM, Chen LC, Hyland SL, Girbes A, Elbers P, Celi LA. 2022. Leveraging electronic health records for data science: common pitfalls and how to avoid them. Lancet Digit Health

56. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. 2018. Promises and pitfalls of electronic health record analysis. Diabetologia 61:1241–8 [PubMed: 29247363]
57. Taksler GB, Dalton JE, Perzynski AT, Rothberg MB, Milinovich A, et al. 2021. Opportunities, Pitfalls, and Alternatives in Adapting Electronic Health Records for Health Services Research. Med Decis Making 41:133–42 [PubMed: 32969760]
58. Callahan A, Shah NH, Chen JH. 2020. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. Ann Intern Med 172:S79–S84 [PubMed: 32479175]
59. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, et al. 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 12:e1001885 [PubMed: 26440803]
60. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, et al. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 147:573–7 [PubMed: 17938396]
61. Baird DD, Dunson DB, Hill MC, Cousins D, Schectman JM. 2003. High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. Am J Obstet. Gynecol 188:100–7 [PubMed: 12548202]
62. Feingold-Link L, Edwards TL, Jones S, Hartmann KE, Velez Edwards DR. 2014. Enhancing uterine fibroid research through utilization of biorepositories linked to electronic medical record data. Journal of women's health (2002) 23:1027–32
63. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12:e1001779 [PubMed: 25826379]
64. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, et al. 2015. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol 44:1137–47 [PubMed: 24518929]
65. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, et al. 2017. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol 27:S2–S8 [PubMed: 28189464]
66. Knudsen LB, Borlum Kristensen F. 1986. Monitoring perinatal mortality and perinatal care with a national register: content and usage of the Danish Medical Birth Register. Community Med 8:29–36 [PubMed: 3698562]
67. Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, et al. 2019. Exome sequencing of Finnish isolates enhances rare-variant association power. Nature 572:323–8 [PubMed: 31367044]
68. Gordis L 2014. Epidemiology Philadelpia, PA: Elsevier Saunders
69. Meinert CL. 2012. Clinical trials: design, conduct, and analysis New York, New York: Oxford University Press
70. Carnes M, Morrissey C, Geller SE. 2008. Women's health and women's leadership in academic medicine: hitting the same glass ceiling? J Womens Health (Larchmt) 17:1453–62 [PubMed: 18954235]
71. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. 2019. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Transactions on Computational Biology and Bioinformatics 16:139–53 [PubMed: 29994486]
72. Wang P, Garza M, Zozus M. 2019. Cancer Phenotype Development: A Literature Review. Stud Health Technol Inform
73. Sharifi-Heris ZA-O, Laitala JA-O, Airola AA-O, Rahmani AA-O, Bender MA-O. 2022. Machine Learning Approach for Preterm Birth Prediction Using Health Records: Systematic Review. JMIR Med Inform 10
74. Sufriyana H, Husnayain A, Chen YL, Kuo CY, Singh O, et al. 2020. Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Med Inform 8:e16503 [PubMed: 33200995]
75. Espinosa C, Becker M, Marić I, Wong RJ, Shaw GM, et al. 2021. Data-Driven Modeling of Pregnancy-Related Complications. Trends Mol Med 27:762–76 [PubMed: 33573911]

76. Bertini A, Salas R, Chabert S, Sobrevia L, Pardo F. 2022. Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review. Front Bioeng Biotechnol
77. Davidson L, Boland MR. 2021. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. Brief Bioinform 22
78. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, et al. 2021. Genome-wide association studies. Nature Reviews Methods Primers 1:59
79. Romualdo Cardoso SA-O, Gillespie A, Haider SA-O, Fletcher OA-O. 2021. Functional annotation of breast cancer risk loci: current progress and future directions. British Journal of Cancer
80. Low SA-O, Zembutsu H, Nakamura Y. 2017. Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Science
81. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47:D1005–D12 [PubMed: 30445434]
82. Mlodawska OW, Saini P, Parker JB, Wei JJ, Bulun SE, et al. 2022. Epigenomic and enhancer dysregulation in uterine leiomyomas. Hum Reprod Update 28:518–47 [PubMed: 35199155]
83. Mohamad MA, Mohd Manzor NF, Zulkifli NF, Zainal N, Hayati AR, Ahmad Asnawi AW. 2020. A Review of Candidate Genes and Pathways in Preeclampsia-An Integrated Bioinformatical Analysis. Biology (Basel) 9
84. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, et al. 2019. The International Cancer Genome Consortium Data Portal. Nat Biotechnol 37:367–9 [PubMed: 30877282]
85. Consortium GT. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369:1318–30 [PubMed: 32913098]
86. Kim G, Jang G, Song J, Kim D, Lee S, et al. 2022. A transcriptome-wide association study of uterine fibroids to identify potential genetic markers and toxic chemicals. PLoS One 17:e0274879 [PubMed: 36174000]
87. Li B, Ritchie MD. 2021. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. Front Genet 12:713230 [PubMed: 34659337]
88. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, et al. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 9:1825 [PubMed: 29739930]
89. Gusev A, Ko A, Shi H, Bhatia G, Chung W, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48:245–52 [PubMed: 26854917]
90. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47:1091–8 [PubMed: 26258848]
91. Edwards TL, Giri A, Hellwege JN, Hartmann KE, Stewart EA, et al. 2019. A Trans-Ethnic Genome-Wide Association Study of Uterine Fibroids. Front Genet 10:511 [PubMed: 31249589]
92. Hellwege JN, Jeff JM, Wise LA, Gallagher CS, Wellons M, et al. 2017. A multi-stage genome-wide association study of uterine fibroids in African Americans. Hum Genet 136:1363–73 [PubMed: 28836065]
93. Chen H, Fan S, Stone J, Thompson DJ, Douglas J, et al. 2022. Genome-wide and transcriptome-wide association studies of mammographic density phenotypes reveal novel loci. Breast Cancer Res 24:27 [PubMed: 35414113]
94. Feng H, Gusev A, Pasaniuc B, Wu L, Long J, et al. 2020. Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. Genet Epidemiol 44:442–68 [PubMed: 32115800]
95. Wu L, Shi W, Long J, Guo X, Michailidou K, et al. 2018. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet 50:968–78 [PubMed: 29915430]
96. Shi J, Wu L, Li B, Lu Y, Guo X, et al. 2019. Transcriptome-Wide Association Study Identifies Susceptibility Loci and Genes for Age at Natural Menopause. Reprod Sci 26:496–502 [PubMed: 29848177]

97. Guintivano J, Aberg KA, Clark SL, Rubinow DR, Sullivan PF, et al. 2022. Transcriptome-wide association study for postpartum depression implicates altered B-cell activation and insulin resistance. Mol Psychiatry 27:2858–67 [PubMed: 35365803]

98. Lu M, Feng R, Qin Y, Deng H, Lian B, et al. 2022. Identifying Environmental Endocrine Disruptors Associated With the Age at Menarche by Integrating a Transcriptome-Wide Association Study With Chemical-Gene-Interaction Analysis. Front Endocrinol (Lausanne) 13:836527 [PubMed: 35282430]

99. Gusev A, Lawrenson K, Lin X, Lyra PC, Jr., Kar S, et al. 2019. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat Genet 51:815–23 [PubMed: 31043753]

100. Lu Y, Beeghly-Fadiel A, Wu L, Guo X, Li B, et al. 2018. A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk. Cancer Res 78:5419–30 [PubMed: 30054336]

101. Zhou T, Huang L, Wang M, Chen D, Chen Z, Jiang SW. 2020. A Critical Review of Proteomic Studies in Gestational Diabetes Mellitus. J Diabetes Res 2020:6450352 [PubMed: 32724825]

102. Penick ER, Bateman NW, Rojas C, Magana C, Conrads K, et al. 2022. Proteomic alterations associated with residual disease in neoadjuvant chemotherapy treated ovarian cancer tissues. Clin Proteomics 19:35 [PubMed: 36195845]

103. Duda JM, Twigg CAI, Thomas SN. 2022. Differential histone deacetylase inhibitor-induced perturbations of the global proteome landscape in the setting of high-grade serous ovarian cancer. Proteomics:e2100372

104. Schon SB, Yang K, Schindler R, Jiang L, Neff LM, et al. 2022. Obesity-related alterations in protein expression in human follicular fluid from women undergoing in-vitro fertilization (IVF). F S Sci

105. Wang X, Zhao M, Guo Z, Song S, Liu S, et al. 2022. Urinary proteomic analysis during pregnancy and its potential application in early prediction of gestational diabetes mellitus and spontaneous abortion. Ann Transl Med 10:736 [PubMed: 35957715]

106. Alesi S, Ghelani D, Mousa A. 2021. Metabolomic Biomarkers in Polycystic Ovary Syndrome: A Review of the Evidence. Semin Reprod Med 39:102–10 [PubMed: 33946122]

107. Craciunas L, Chu J, Pickering O, Mohiyiddeen L, Coomarasamy A. 2022. The metabolomic profile of endometrial receptivity in recurrent miscarriage. Minerva Obstet Gynecol

108. Troisi J, Mollo A, Lombardi M, Scala G, Richards SM, et al. 2022. The Metabolomic Approach for the Screening of Endometrial Cancer: Validation from a Large Cohort of Women Scheduled for Gynecological Surgery. Biomolecules 12

109. Yu X, Chen Y, He L, Liu H, Yang Z, Lin Y. 2022. Transcriptome and metabolome analyses reveal the interweaving of immune response and metabolic regulation in pelvic organ prolapse. Int Urogynecol J

110. Ahmeidat A, Kotts WJ, Wong J, McLernon DJ, Black M. 2021. Predictive models of individual risk of elective caesarean section complications: a systematic review. Eur J Obstet Gynecol Reprod Biol 262:248–55 [PubMed: 34090730]

111. Black N, Henderson I, Al Wattar BH, Quenby S. 2022. Predictive Models for Estimating the Probability of Successful Vaginal Birth After Cesarean Delivery: A Systematic Review. Obstet Gynecol

112. Deng B, Li Y, Chen JY, Guo J, Tan J, et al. 2022. Prediction models of vaginal birth after cesarean delivery: A systematic review. Int J Nurs Stud 135:104359 [PubMed: 36152466]

113. Mazo C, Aura C, Rahman A, Gallagher WM, Mooney C. 2022. Application of Artificial Intelligence Techniques to Predict Risk of Recurrence of Breast Cancer: A Systematic Review. J Pers Med 12

114. Zhong M, Zhang H, Yu C, Jiang J, Duan X. 2022. Application of machine learning in predicting the risk of postpartum depression: A systematic review. J Affect Disord 318:364–79 [PubMed: 36055532]

115. Carr BL, Jahangirifar M, Nicholson AE, Li W, Mol BW, Licqurish S. 2022. Predicting postpartum haemorrhage: A systematic review of prognostic models. Aust N Z J Obstet Gynaecol

116. Raeisi-Dehkordi H, Kummer S, Francis Raguindin P, Dejanovic G, Eylul Taneri P, et al. 2022. Risk Prediction Models of Natural Menopause Onset: A Systematic Review. J Clin Endocrinol Metab 107:2934–44 [PubMed: 35908226]
117. Joo YY, Actkins K, Pacheco JA, Basile AO, Carroll R, et al. A Polygenic and Phenotypic Risk Prediction for Polycystic Ovary Syndrome Evaluated by Phenome-Wide Association Studies
118. Svensson AA-O, Garcia-Etxebarria K, Åkesson A, Borgfeldt C, Roth B, et al. Applicability of polygenic risk scores in endometriosis clinical presentation
119. Piekos JA-O, Hellwege JN, Zhang Y, Torstenson ES, Jarvik GP, et al. Uterine fibroid polygenic risk score (PRS) associates and predicts risk for uterine fibroid. LID - 10.1007/s00439-022-02442-z [doi].
120. Dareng EA-OX, Tyrer JA-O, Barnes DA-O, Jones MR, Yang XA-O, et al. Polygenic risk modeling for prediction of epithelial ovarian cancer risk
121. Yanes TA-O, Young MA, Meiser B, James PA. 2020. Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. Breast Cancer Research 22
122. Lee A, Mavaddat N, Cunningham A, Carver T, Ficorella L, et al. 2022. Enhancing the BOADICEA cancer risk prediction model to incorporate new data on RAD51C, RAD51D, BARD1 updates to tumour pathology and cancer incidence. J Med Genet
123. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, et al. 2019. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med 21:1708–18 [PubMed: 30643217]
124. Yang X, Eriksson M, Czene K, Lee A, Leslie G, et al. 2022. Prospective validation of the BOADICEA multifactorial breast cancer risk prediction model in a large prospective cohort study. J Med Genet
125. Rafnar T, Gunnarsson B, Stefansson OA, Sulem P, Ingason A, et al. 2018. Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. Nat Commun 9:3636 [PubMed: 30194396]
126. Day F, Karaderi T, Jones MR, Meun C, He C, et al. 2018. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. PLoS Genet 14:e1007813 [PubMed: 30566500]
127. Adewuyi EO, Mehta D, International Endogene C, andMe Research T, Nyholt DR. 2022. Genetic overlap analysis of endometriosis and asthma identifies shared loci implicating sex hormones and thyroid signalling pathways. Hum Reprod 37:366–83 [PubMed: 35472084]
128. Olafsdottir T, Thorleifsson G, Sulem P, Stefansson OA, Medek H, et al. 2020. Genome-wide association identifies seven loci for pelvic organ prolapse in Iceland and the UK Biobank. Commun Biol 3:129 [PubMed: 32184442]
129. Steinthorsdottir V, McGinnis R, Williams NO, Stefansdottir L, Thorleifsson G, et al. 2020. Genetic predisposition to hypertension is associated with preeclampsia in European and Central Asian women. Nat Commun 11:5976 [PubMed: 33239696]
130. Zhang G, Feenstra B, Bacelis J, Liu X, Muglia LM, et al. 2017. Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. N Engl J Med 377:1156–67 [PubMed: 28877031]
131. Mirin AA. 2021. Gender disparity in the funding of diseases by the US National Institutes of Health. Journal of Women's Health 30:956–63
132. King Thomas J, Mir H, Kapur N, Singh S. 2019. Racial differences in immunological landscape modifiers contributing to disparity in prostate cancer. Cancers 11:1857 [PubMed: 31769418]
133. Oliveira DF, Ma Y, Woodruff TK, Uzzi B. 2019. Comparison of National Institutes of Health grant amounts to first-time male and female principal investigators. Jama 321:898–900 [PubMed: 30835300]
134. Hechtman LA, Moore NP, Schulkey CE, Miklos AC, Calcagno AM, et al. 2018. NIH funding longevity by gender. Proceedings of the National Academy of Sciences 115:7943–8
135. Knobloch-Westerwick S, Glynn CJ, Huge M. 2013. The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. Science communication 35:603–25

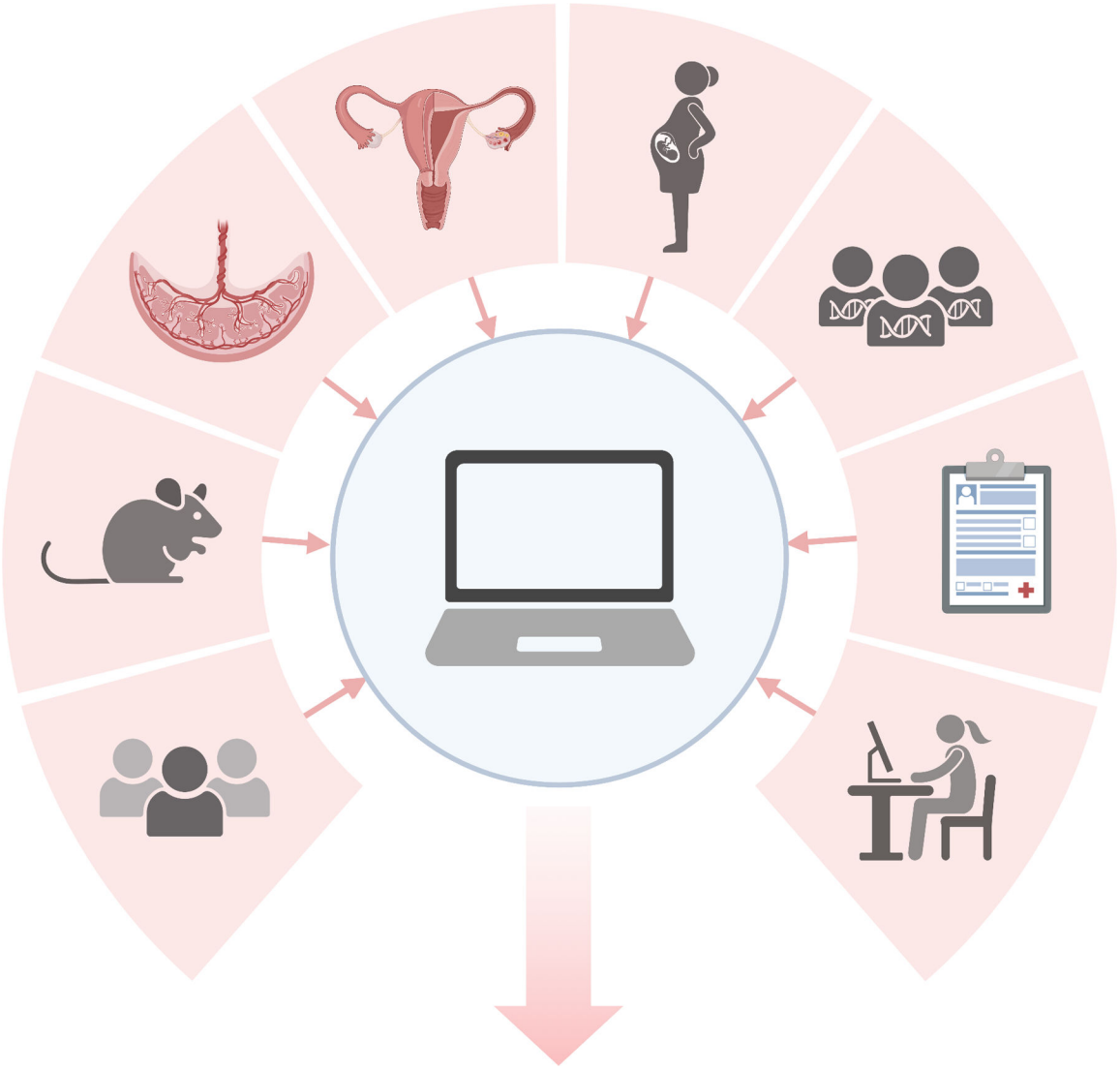


Figure 1.
An overview of the domains of research that encompass women's health and data science. Created with BioRender.com.

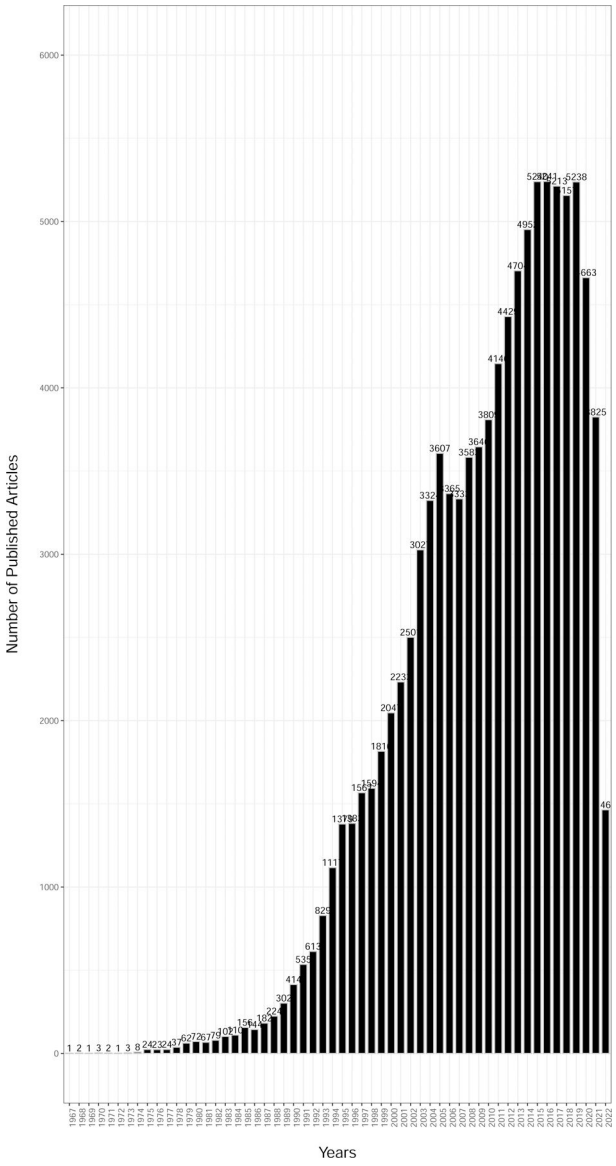
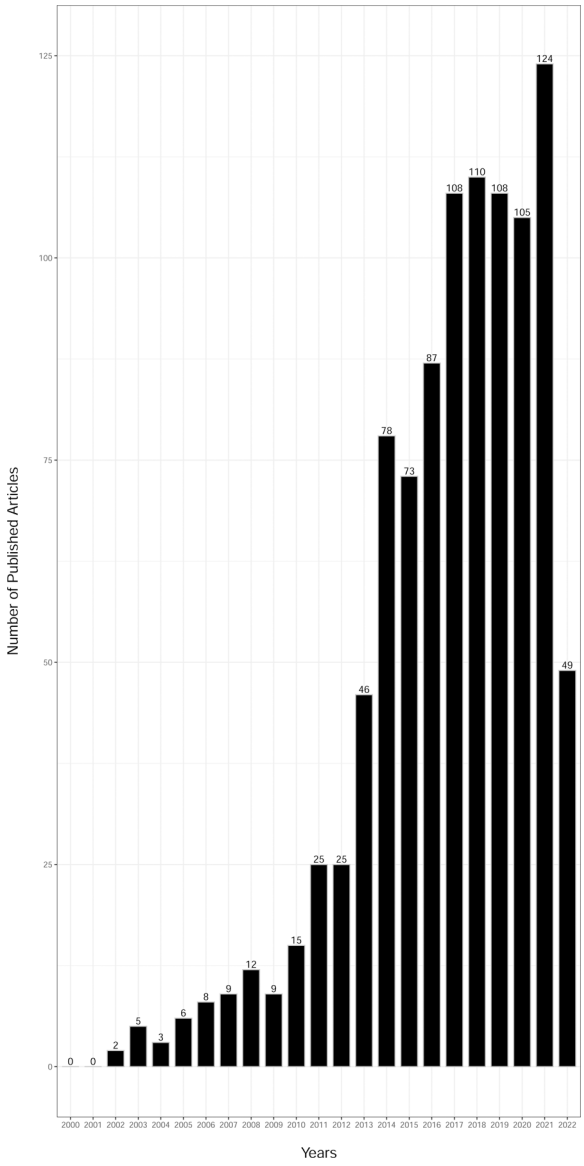


Figure 2.
Published literature using data science approaches from women's health research (1967–2022) (n = 101,623).



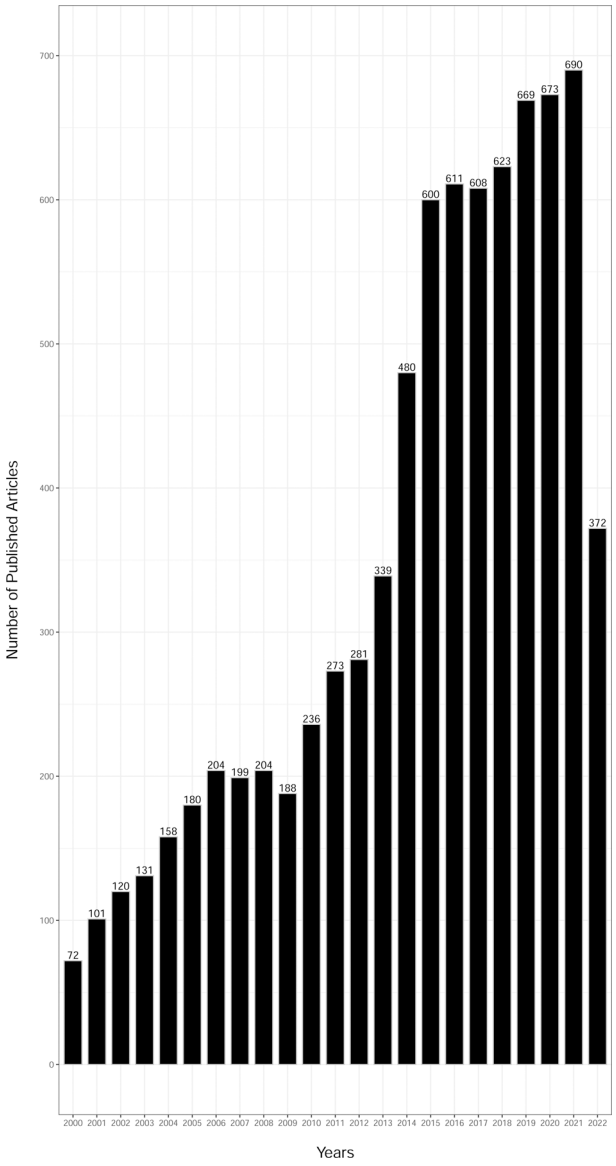


Figure 3.

Published literature (2000–2022) from women's health research using data science approaches and (*a*) electronic health records or (*b*) genomics.

Annu Rev Biomed Data Sci. Author manuscript; available in PMC 2024 August 10.

Table 1.

Observational Population Research Study Designs

	Cohort Studies		
	Prospective	**Retrospective**	**Case-Control Studies**
Study Group	Exposed persons (a+b)	Exposed persons (a+b)	Persons with the disease (cases): (a+c)
Comparison Group	Nonexposed persons (c+d)	Nonexposed persons (c+d)	Persons without disease (controls): (b+d)
Outcome Measurements	Incidence in the exposed $\frac{a}{(a+b)}$ And incidence in the nonexposed $\frac{c}{(c+d)}$	Incidence in the exposed $\frac{a}{(a+b)}$ And incidence in the nonexposed $\frac{c}{(c+d)}$	Proportion of cases exposed $\frac{a}{(a+c)}$ And proportion of controls exposed $\frac{b}{(b+d)}$
Measures of Risk	Absolute Risk, Relative Risk, Odds Ratio, Attributable Risk	Absolute Risk, Relative Risk, Odds Ratio, Attributable Risk	Odds ratio and attributable Risk.
Temporal Relationship between exposure and disease	Easy to establish	Sometimes hard to establish	Sometimes hard to establish
Multiple Associations	Possible to study association of an exposure with several diseases	Possible to study association of an exposure with several diseases	Possible to study association of an exposure with several diseases
Time required for the study	Generally, long because of need to follow-up the subjects	May be short	Relatively short
Cost of study	Expensive	Generally, less expensive than prospective studies	Relatively inexpensive
Population size needed	Relatively large	Relatively large	Relatively small
Potential Bias	Assessment of outcome	Susceptible to bias both in assessment of exposure and assessment of outcome	Assessment of exposure
Best when	Exposure is rare, disease is frequent among exposed	Exposure is rare, disease is frequent among exposed	Disease is rare and exposure is frequent among persons with disease
Problems	Selection of nonexposed comparison group often difficult. Changes over time in criteria and methods.	Selection of nonexposed comparison group often difficult. Changes over time in criteria and methods.	Selection of appropriate controls often difficult. Incomplete information on exposure.