Of Mice and Women

The Health Returns to Sex-Inclusive Biomedical Research in the US

Extended Abstract
Preliminary — do not circulate

Ottavia Capozzi
Universitat Pompeu Fabra

November 2, 2026

Research question

For decades biomedical research was conducted predominantly on male models, producing a scientific literature whose clinical translation is systematically less reliable for women. Until 1993 women of childbearing potential were routinely excluded from US clinical trials, and the asymmetry persisted far longer in preclinical science: Beery and Zucker (2011) document a male bias in 8 of 10 biological fields, with male-only studies outnumbering female-only studies by 5.5 to 1 in neuroscience. Because pharmacokinetics, symptom presentation, disease progression and drug metabolism differ systematically between the sexes, evidence generated on male models translates less reliably to female patients (Prendergast et al., 2014), and the documented downstream consequences are a disproportionate female burden of adverse drug reactions, diagnostic delay and suboptimal dosing (Zucker and Prendergast, 2020; Franconi and Campesi, 2017).

This paper asks whether the wave of sex-inclusivity mandates that reached US biomedical research in the mid-2010s actually changed what researchers do, and it sets up the chain through which that change would reach patients. The question matters for economics because the mandates are an unusually clean case of regulation acting on the *production technology of knowledge*: they do not fund more research, they prescribe how research must be designed, analysed and reported. Whether such prescriptions bind—and whether they bind on the margin that matters for clinical translation—is an open empirical question.

Policy setting

I study three overlapping mandates that reached the research pipeline within roughly two years of each other, through three distinct channels.

1. **NIH Sex as a Biological Variable (SABV)** — the *funder* channel. Announced in May 2014 and codified in Notice NOT-OD-15-102, released 9 June 2015 and effective for applications submitted on or after 25 January 2016, SABV requires NIH-funded researchers to factor sex into the design, analysis and reporting of vertebrate animal and human studies, and to justify single-sex designs (National Institutes of Health, 2015; Clayton and Collins, 2014).

2. **The 21st Century Cures Act** — the *registry* channel. Enacted 13 December 2016, it requires applicable NIH-defined Phase III clinical trials to report results of valid analyses

by sex/gender and race/ethnicity to ClinicalTrials.gov; NIH implemented the requirement through Notice NOT-OD-18-014 for competing awards made on or after 13 December 2017.

3. **The SAGER guidelines** — the *journal* channel. Published in 2016 by the European Association of Science Editors' Gender Policy Committee, the Sex and Gender Equity in Research guidelines ask journals to require that authors report the sex of animals, cells and participants and, where the sample permits, present results disaggregated by sex (Heidari et al., 2016). SAGER binds at the point of publication rather than the point of award, and it is not restricted to NIH-funded work.

Including SAGER is essential to the design rather than incidental to it. SABV operates on NIH grantees and the Cures Act on registered Phase III trials, so a naive NIH-versus-non-NIH comparison implicitly assumes that non-NIH research is untreated. SAGER breaks that assumption: an author submitting to a SAGER-endorsing journal faces a sex-reporting requirement whatever the source of their funding. Together with the FDAAA Final Rule (42 CFR Part 11, effective January 2017), which changed ClinicalTrials.gov reporting obligations for non-NIH trials as well, this means the control group is itself partially treated—a violation of SUTVA that motivates the second of the two designs below.

Data

I build a corpus of biomedical articles published 2008–2025 from PubMed Central, whose full text makes the reporting outcomes observable rather than inferred. For each article I extract the publication year, journal, MeSH terms, medical specialty (WHO classification), CCSR medical condition, author count and share of female authors, and I identify NIH funding from grant acknowledgements. Sex-reporting outcomes are extracted from the abstract, methods, results and discussion of each paper, and coded as four mutually exclusive descriptions of *which* subjects were studied—only male, only female, both sexes, sex not reported—plus a separate indicator for whether the paper went on to *analyse* its results separately by sex. The corpus contains $N = 294{,}376$ papers, of which 194,277 are clinical and 100,099 preclinical. The two layers are worth distinguishing: inclusion is a design choice, sex-differentiated analysis is an interpretive one, and the second is far rarer than the first. Among preclinical papers that do break results down by sex, 87.8% report finding a difference and 12.2% report finding none (88.1% against 11.9% in the clinical corpus), so the analysis indicator records the act of looking, which in practice almost always coincides with reporting a difference.

The downstream stages draw on four further sources. **Clinical practice guidelines** are retrieved from PubMed and coded in full text on the same sex-reporting dimensions as the research corpus, which makes guideline documents directly comparable to the literature they cite. **Drug labels** come from DailyMed (2010–2025), which archives label revisions and therefore allows sex-related label content to be tracked over time rather than observed only at approval; Drugs@FDA and the NDC directories supply the application type (NDA, ANDA, BLA) and the approval information—approval date, marketing status and withdrawal date. Restricting attention to NDA originators gives an estimation sample of 25,421 drug-year observations covering 2,329 drugs, 138 medical conditions and 21 specialties, which is the panel on which the label analysis is run. Finally, the **Medical Expenditure Panel Survey** (MEPS 2010–2023, adults 18–64) provides expenditures, utilisation and self-reported outcomes, and the **FDA Adverse Event Reporting System** (FAERS 2010–2023) provides sex-stratified adverse event counts. The condition coding is common to all of them, so the dose B_c defined below carries across papers, guidelines, labels and patients.

Empirical strategy

The unit of observation is a paper i in year t; the data are repeated cross-sections.

Design 1: binary difference-in-differences. Where a credible control group exists—the preclinical corpus, which the Cures Act does not reach—I compare NIH-funded papers ($D_i = 1$) with non-NIH-funded papers in the same specialties and window ($D_i = 0$), estimating by `reghdfe`

$$Y_{i,t} = \alpha_{s(i)} + \delta_{j(i)} + \gamma_t + \lambda_{s(i)}\, t + \mu_{j(i)}\, t + \sum_{\substack{e=-7 \\ e\neq -3}}^{9} \beta_e \mathbb{1}[\,t - 2016 = e\,]\, D_i + X'_{i,t}\theta + u_{i,t}, \tag{1}$$

with medical-specialty, journal and year fixed effects ($\alpha_{s(i)}$, $\delta_{j(i)}$, γ_t), specialty- and journal-specific linear trends, and $X_{i,t}$ collecting author count, the share of female authors, NIH RCDC funding for the paper's disease area and that area's share of PubMed output. The reference period is $e = -3$ (2013) rather than $e = -1$, so that anticipation of the May 2014 announcement is not differenced away. Errors are clustered on the medical specialty.

Design 2: continuous difference-in-differences. For the clinical corpus equation (1) is not identified: the Cures Act, the FDAAA rule and SAGER reach both groups. I replace the binary treatment with a dose that varies across conditions and treat both groups as exposed:

$$Y_{i,c,t} = \alpha_{s(i)} + \delta_{j(i)} + \gamma_t + \beta_0 B_c + \sum_{\substack{e=-7 \\ e\neq -3}}^{9} \beta_e \mathbb{1}[\,t - 2016 = e\,]\, B_c + \phi\, D_i + X'_{i,c,t}\theta + u_{i,c,t}. \tag{2}$$

B_c is the pre-policy *sex-differentiated-analysis (SDA) gap* of condition c; D_i is retained as a control so that β_e does not simply recover the binary funding channel; $X_{i,c,t}$ adds the field's share of female authors and the number of drugs the paper tests. Specialty and journal trends are dropped—the dose varies only across conditions, so such trends would absorb the identifying variation. Errors are clustered on the condition.

The dose. With $\mathcal{I}_c$ the papers in CCSR condition c over $2008 \le t \le 2013$ and $\mathbb{1}[\text{sda}_i]$ indicating a sex-differentiated analysis, the raw gap is one minus the baseline share,

$$\widetilde{B}_c \;=\; 1 \;-\; \frac{\sum_{i\in\mathcal{I}_c} \mathbb{1}[\text{sda}_i]}{|\mathcal{I}_c|}, \tag{3}$$

rescaled across conditions to

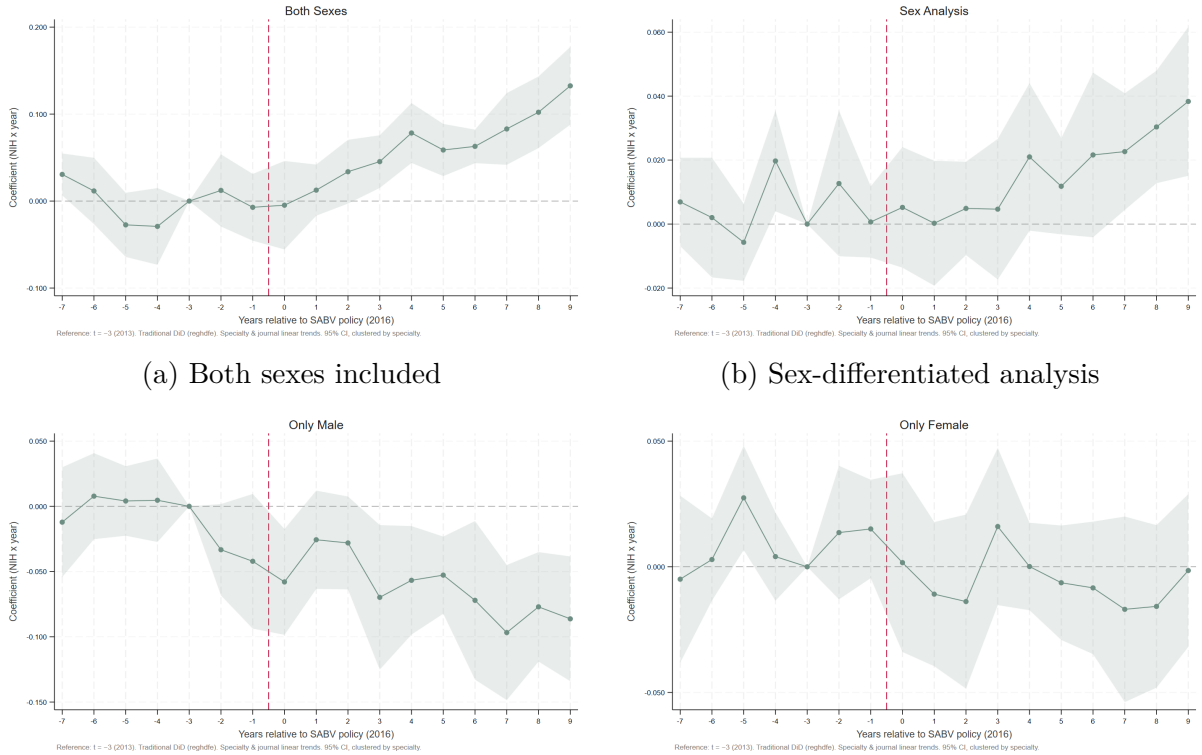
$$B_c \;=\; \frac{\widetilde{B}_c - \min_{c'} \widetilde{B}_{c'}}{\max_{c'} \widetilde{B}_{c'} - \min_{c'} \widetilde{B}_{c'}} \;\in\; [0,1]. \tag{4}$$

Conditions with fewer than 50 baseline papers are dropped. $B_c = 1$ is the condition whose baseline literature analysed its results by sex least often—the largest effective dose. Built from pre-policy behaviour alone, $\widetilde{B}_c$ is fixed before either mandate takes effect; in the clinical corpus it ranges from 0.748 to 1.000, so even the best-performing conditions analysed fewer than one paper in four by sex before 2014.

Results

Preclinical: the binary design. Equation (1) on the preclinical corpus ($N = 73{,}566$; 67 specialty clusters) shows that the mandate moved the *composition* of preclinical science. Both-sex inclusion becomes significant at $e = 3$ (+0.045, $p = 0.004$) and reaches +13.2 percentage points by $e = 9$ ($p < 0.001$); only-male studies fall 8.6 points ($p = 0.001$) and are already down in the implementation year (-0.058, $p = 0.006$); only-female reporting is flat throughout (-0.001 at $e = 9$, $p = 0.92$). Pre-periods show no systematic trend, though only-male reporting drifts down at $e = -2$ and $e = -1$ ($p = 0.06$ and $p = 0.11$), consistent with anticipation of the May 2014 announcement.

The interpretive margin lags badly. Sex-differentiated analysis is indistinguishable from zero until $e = 7$ and reaches only +3.8 points by $e = 9$ ($p = 0.002$)—under a third of the inclusion effect. Papers include both sexes years before they analyse what those sexes do differently, which is the margin on which a difference can actually be discovered.



(a) Both sexes included

(b) Sex-differentiated analysis

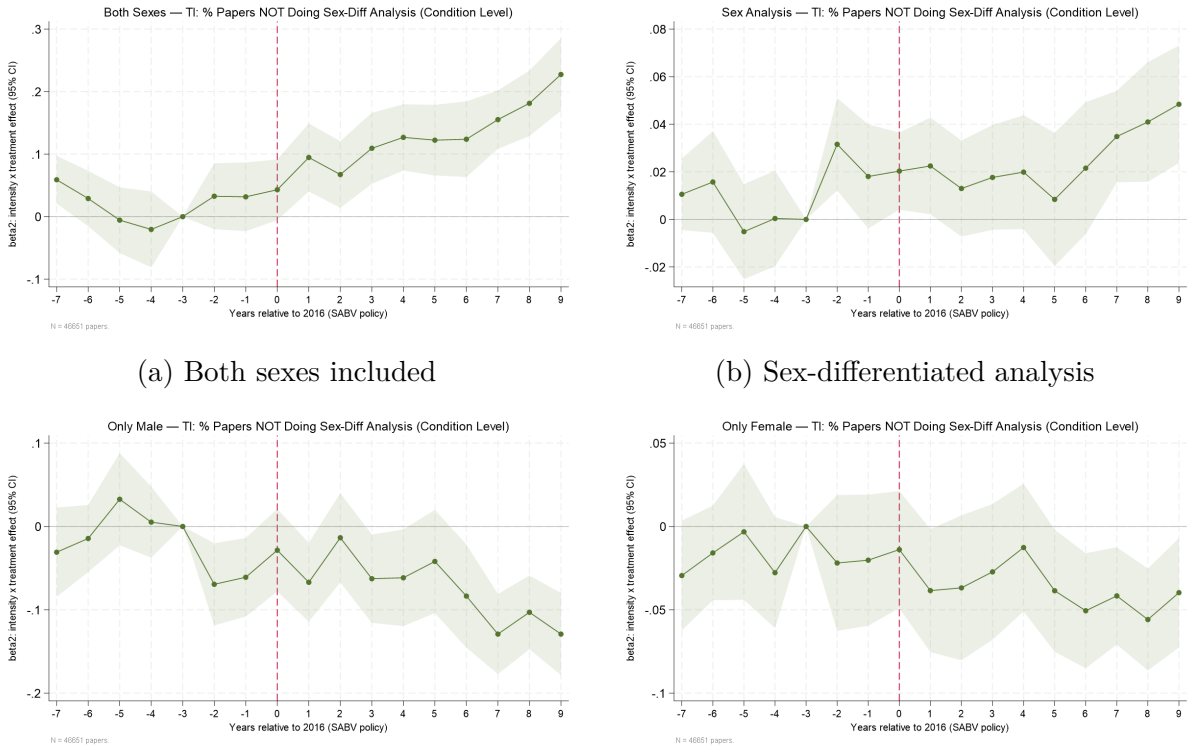
(c) Only male

(d) Only female

Figure 1: **Preclinical papers: binary difference-in-differences.** Estimates of β_e from equation (1), NIH- versus non-NIH-funded papers ($N = 73{,}566$). Year, specialty and journal fixed effects plus specialty- and journal-specific linear trends; reference period 2013 ($e = -3$); 95% CI clustered on the specialty; dashed line marks the 2016 implementation. Inclusion (a) rises roughly three and a half times as much as analysis (b), and the gain comes out of male-only studies (c) rather than female-only ones (d).

Preclinical: the dose design. Replacing the binary treatment with the condition-level dose of equations (3)–(4) on the same corpus ($N = 46{,}651$; 333 condition clusters) sharpens every margin. Per unit of dose, both-sex inclusion rises +22.7 points by $e = 9$ and only-male reporting falls

12.9 points (both $p < 0.001$); only-female reporting declines mildly (-4.0 points, $p = 0.018$); and sex-differentiated analysis rises $+4.8$ points ($p < 0.001$), again the smallest of the four and a fifth of the inclusion effect. Pre-trends are flat for both-sex inclusion but not elsewhere: only-male reporting is already significantly down at $e = -2$ and $e = -1$ (-0.069 and -0.061, $p < 0.05$) and sex-differentiated analysis is already up at $e = -2$ ($+0.032$, $p = 0.001$)—consistent with anticipation of the May 2014 announcement, but a caveat for parallel trends on those two margins. The two preclinical designs agree on the ordering: the conditions with furthest to go moved most, and inclusion moved far more than analysis in both.



(a) Both sexes included

(b) Sex-differentiated analysis

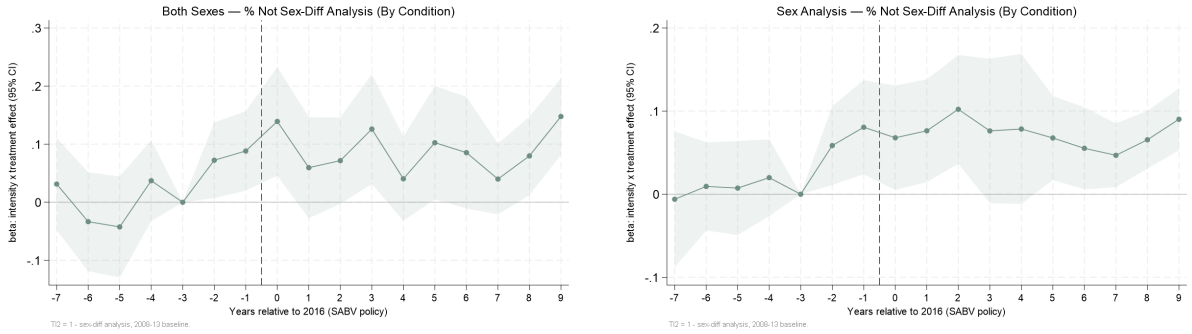
(c) Only male

(d) Only female

Figure 2: **Preclinical papers: continuous difference-in-differences.** Dose-response estimates of β_e from equation (2) on the preclinical corpus, with the condition-level SDA gap as treatment intensity (N = 46,651; 333 CCSR condition clusters). Coefficients are per unit of dose. Year, pooled specialty and journal fixed effects; reference period 2013 ($e = -3$); 95% CI clustered on the condition.

Clinical: the continuous design. Equation (2) on the clinical corpus (N = 109,907; 116 CCSR condition clusters) gives three findings. First, the response dates to the *announcement*, not the implementation: both-sex inclusion is already $+7.2$ and $+8.8$ points per unit of dose at $e = -2$ and $e = -1$ ($p = 0.029$, $p = 0.012$), sex-differentiated analysis $+5.9$ and $+8.1$ points ($p = 0.016$, $p = 0.006$). Differencing against 2015 would net out most of the effect, which is why the reference period is 2013. Second, both margins move, and the interpretive one proportionately more than preclinically: inclusion jumps $+13.9$ points in the implementation year ($p = 0.004$) and ends at $+14.8$ ($p < 0.001$), while sex-differentiated analysis runs $+4.7$ to $+10.2$ across the post-period and ends at $+9.0$ points ($p < 0.001$). Conditions whose literature had furthest to go are the ones that

moved. Third, the single-sex margins are untouched (+0.020, $p = 0.201$ for only-male; −0.006, $p = 0.827$ for only-female), since single-sex designs were already rare in trials: the clinical effect runs entirely through inclusion and analysis.



(a) Both sexes included

(b) Sex-differentiated analysis

Figure 3: **Clinical papers: continuous difference-in-differences.** Dose-response estimates of β_e from equation (2), with the pre-policy SDA gap of equations (3)–(4) as treatment intensity (N = 109,907; 116 CCSR condition clusters). Coefficients are per unit of dose: the difference between a condition that never analysed its results by sex at baseline and one that did so most often. Year, pooled specialty and journal fixed effects; controls as in the text; reference period 2013 ($e = -3$); 95% CI clustered on the condition. Both series lift off at $e = -2$, in the year of the announcement.

Across the three designs the mandates bound, they bound hardest where baseline practice was worst, and they bound earlier and harder on inclusion than on analysis.

From papers to patients: work in progress

The results above are the first stage of the project. What makes the question an economic one is whether the change in research practice propagates through the institutions that convert evidence into treatment. Three stages follow, and each is a separate empirical exercise.

(i) Clinical practice guidelines. Guidelines are the proximate input into physician behaviour, and they are written by committees that read the literature I have just measured. The guideline corpus is in hand, and the coding asks, document by document and revision by revision, whether recommendations are sex-specific—distinct dosing, screening ages, risk thresholds or diagnostic criteria for women—and whether the supporting evidence cited is itself sex-disaggregated. Mapping guidelines to CCSR conditions makes the dose B_c of equation (4) directly applicable, so the same event-study design can be run on guideline revisions, with the lag structure of guideline updating as the object of interest.

(ii) FDA drug labels. Labels are where sex-specific evidence becomes legally operative: the Clinical Pharmacology, Dosage and Administration, and Use in Specific Populations sections are what a prescriber reads. Because DailyMed archives revisions, the label panel is a genuine time series rather than a snapshot at approval: I link the drug-testing subsamples of both corpora to it through the compounds each paper studies, and ask whether drugs whose underlying literature became more sex-differentiated acquire more sex-specific label content—and, since Cleary et al.

(2023) show that NIH funding contributed to published research associated with 354 of the 356 new drugs approved between 2010 and 2019, this is a channel through which a research mandate can reach prescribing at scale. FAERS provides the complementary safety test: if labels improve, the female excess in sex-stratified adverse event rates should narrow for exactly those drugs.

(iii) Women's outcomes and expenditures. The final stage asks what any of this is worth. Using MEPS, I estimate whether women in high-dose conditions—those whose research literature changed most—experience differential post-2016 changes in healthcare expenditure, utilisation, cost-related access barriers and self-reported health relative to men in the same conditions and to women in low-dose conditions. The expenditure margin is substantial at baseline: women's per-capita spending exceeds men's across the adult lifespan (Lassman et al., 2014; Cylus et al., 2011), the gap persists within narrowly defined conditions such as diabetes (Williams et al., 2017), and women spent over $8.5 billion more than men on prescription drugs in a single year (Marsh, 2025). Whether better science reduces that gap, and on which margin, is the question the project is ultimately built to answer.

References

E. Andrew Balas and Stephen A. Boren. Managing clinical knowledge for health care improvement. *Yearbook of Medical Informatics*, pages 65–70, 2000. Documents 17-year average lag from medical discovery to clinical adoption.

Annaliese K. Beery and Irving Zucker. Sex bias in neuroscience and biomedical research. *Neuroscience and Biobehavioral Reviews*, 35(3):565–572, 2011.

Janine A. Clayton and Francis S. Collins. Policy: Nih to balance sex in cell and animal studies. *Nature*, 509 (7500):282–283, 2014.

Ekaterina Galkina Cleary, Matthew J. Jackson, Edward W. Zhou, and Fred D. Ledley. Comparison of research spending on new drug approvals by the national institutes of health vs the pharmaceutical industry, 2010–2019. *JAMA Health Forum*, 4(4):e230511, 2023. doi: 10.1001/jamahealthforum.2023.0511.

Jonathan Cylus, Micah Hartman, Benjamin Washington, Kimberly Andrews, and Aaron Catlin. Pronounced gender and age differences are evident in personal health care spending per person. *Health Affairs*, 30(1): 153–160, 2011.

Doximity. Survey: How doctors read and what it means to patients. Business Wire, July 2014. URL https://www.businesswire.com/news/home/20140722005535/en/. Accessed: 2026.

Flavia Franconi and Isabella Campesi. Sex impact on biomarkers, pharmacokinetics and pharmacodynamics. *Current Medicinal Chemistry*, 24(24):2561–2575, 2017. doi: 10.2174/0929867323666161003124616.

Shirin Heidari, Thomas F. Babor, Paola De Castro, Sera Tort, and Mirjam Curno. Sex and gender equity in research: rationale for the sager guidelines and recommended use. *Research Integrity and Peer Review*, 1:2, 2016. doi: 10.1186/s41073-016-0007-6.

David Lassman, Micah Hartman, Benjamin Washington, Kimberly Andrews, and Aaron Catlin. Us health spending trends by age and gender: Selected years 2002–10. *Health Affairs*, 33(5):815–822, 2014.

Tori Marsh. The prescription drug gender divide: Women spent over $8.5 billion more than men in 2024. GoodRx Research, March 2025. URL https://www.goodrx.com/healthcare-access/research/prescription-drug-gender-gap-women-spend-more. March 7, 2025.

National Institutes of Health. Consideration of sex as a biological variable in nih-funded research, 2015. Notice NOT-OD-15-102, Office of Extramural Research.

Brian J. Prendergast, Kenneth G. Onishi, and Irving Zucker. Female mice liberated for inclusion in neuroscience and biomedical research. *Neuroscience and Biobehavioral Reviews*, 40:1–5, 2014. doi: 10.1016/j.neubiorev.2014.01.001.

Joni S. Williams, Kinfe Bishu, Clara E. Dismuke, and Leonard E. Egede. Sex differences in healthcare expenditures among adults with diabetes: Evidence from the medical expenditure panel survey, 2002–2011. *BMC Health Services Research*, 17, 2017. doi: 10.1186/s12913-017-2178-3.

Irving Zucker and Brian J. Prendergast. Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biology of Sex Differences*, 11:32, 2020. doi: 10.1186/s13293-020-00308-5.