

科学人才流失模式: OECD 国家研究格局的定量分析及方法论考量*

马瑞克·科维克 (Marek Kwiek)¹ 卢卡斯·希穆拉 (Lukasz Szymula)²

(1. 波兰波兹南大学公共政策研究中心, 波兹南 61712;
2. 波兰波兹南大学数学与计算机科学学院, 波兹南 61712)

摘要: 采用基于队列的纵向研究设计, 借助 Scopus 全球文献计量数据库, 追踪分析了 38 个 OECD 国家科研人员截至 2022 年的科研发表数据, 探讨了科学界成员是如何离开学术界的, 以及科学人才流失在性别、学科领域和时间上的差异。研究涉及 2000 年 (N=142,776) 和 2010 年 (N=232,843) 开始发表论文的两批科研人员, 覆盖科学、技术、工程、数学与医学 (STEMM) 领域的 16 个学科, 以停止发表论文作为科研人员离开学术界的界定标准。研究显示: 初次发表论文十年后, 约 50% 的研究者仍保持学术活跃状态; 当追踪周期延长至十九年后, 该比例降至 30%。通过生存分析法比较男女科学家的流失差异发现, 随着女性在科学领域及同批次科研人员中占比的增加, 人才流失的性别差异逐渐弱化。学科维度分析揭示, 除了所有 STEMM 领域的整体变化外, 具体学科层面的细微变化广泛存在。不同学科在科学人才流失上呈现出不同的性别差异; 不同时间进入科学界的科研人员在流失概率上也存在差别。本研究验证了全球文献计量数据库在分析科学人才流失中的应用价值, 并指出了原始结构化数据在学术职业的性别、年龄和学科等研究中的方法挑战和局限性。

关键词: 学术生涯; 科研流失; 大数据; 生存分析; 科学学

按语: 马瑞克·科维克 (Marek Kwiek), 波兰波兹南 AMU 大学社会科学与人文学高等研究院 (IAS) 主任, 教授、联合国教科文组织机构研究与高等教育政策教席主席、公共政策研究中心 (CPPS) 主任。他的研究领域为科学量化研究、科学社会学及高等教育研究, 目前主要关注国际研究合作、学术生产力、科学阶层分化及全球学术精英等。该团队的系列相关研究是迄今量化离开科学界人数的最大规模研究, 并于 2024 年和 2025 年被《自然》报道。

本研究采用队列研究 (Glenn, 2005) 与纵向追踪方法 (Singer & Willett, 2003), 以停止发表论文为科学人才离开学术界的界定标准, 系统分析科学人才流失率在性别、学科及时序维度上的差异。鉴于全球范围内缺乏学术职业离开行为的大规模纵向观测数据, 本研究对男性和女性科学家进行长期跟踪, 并对传统上被称为“离开科学界”的现象进行量化分析 (Geuna & Shibayama, 2015; Preston, 2004; White-Lewis et al., 2023; Zhou & Volkwein, 2004)。利用 Scopus 中的文献计量出版元数据, 追踪了来自 38 个经合组织国家的发表记录分别始于 2000 年和 2010 年的 142,776 名和 232,843 名科学家的科研发表记录 (分别称为 2000 年和 2010 年队列)。研究仅限于 16 个 STEMM 学科 (科学、技术、工程、数学和医学), 两批科学家的学术产出成果统计至 2022 年。

文献计量元数据作为一种数字痕迹, 突破了传统的调查、访谈及行政数据在覆盖范围与时间跨度

* 基金项目: 波兰教育和科学部科学协会资助项目“大数据如何丰富学术专业研究? 从理论模型到实际应用” (NdS-II/SP/0010/2023/01)

上的局限,推动了科学家群体研究范式的革新(Kashyap et al., 2023; Liu et al., 2023; Wang & Barabási, 2021)。这种数据形态的转变使研究者得以将分析单元从单篇论文扩展至科学家个体层面,通过对学术产出轨迹的持续追踪,系统考察年龄、资历、性别、学科归属及机构类型等多维度变量的交互影响。此外,横断面研究可以通过纵向研究来补充,纵向研究可以跟踪个人的学术生涯数年甚至数十年。利用海量数据集,研究人员在全球和国家层面对学术生涯进行了多项研究(Ioannidis et al., 2014; King et al., 2017; Larivière et al. 2013; Milojevic et al., 2018; Nielsen & Andersen, 2021; Nygaard et al., 2022; Spoon et al., 2023)。

既往研究表明,科学人才流失存在显著的性别差异:女性科研人员不仅较男性更早中断学术生涯,其总体流失比例亦高于男性,这一现象与高等教育领域人员流失的性别差异的经典论述形成呼应(Blickenstaff, 2005; Deutsch & Yao, 2014; Goulden et al., 2011; Preston, 2004; Shaw & Stanton, 2012)。本研究在方法论上实现双重突破:其一,突破传统横断面研究对瞬时状态(单次或重复时间快照)的依赖,通过 38 个 OECD 国家科学家个体层面的纵向追踪数据,系统刻画学术出版生涯中的流失动态;其二,创新性验证全球文献计量数据集在大规模量化学术职业中断现象中的可行性。这种数据驱动的研究范式拓展了科技人力资源研究的时空尺度,为解构科研界人员流失的性别差异提供了新的实证基础。

一、理论背景: 男性科学家、女性科学家与科学界的人才流失

不论男性还是女性科学家在进入和选择学术生涯时均面临多重挑战(Preston, 2004; Wohrer, 2014)。然而,离开学术界历来被视为女性主导的现象:有研究声称,女性在学术界面临着“寒冷”的工作场所文化、难以维持工作与生活平衡以及在学术工作中难以度过生育期等问题(Cornelius et al., 1988; Goulden et al., 2011; Maranto & Griffin, 2011; White-Lewis et al., 2023)。这些障碍主要体现在学术晋升、科研产出、学术影响力、科研经费获取、奖项争取及学术成就认可等方面。如女性在高级学术职位中代表性不足;在国际科研合作、高影响力期刊发文及高被引概率方面处于劣势;更易遭遇较长的职业中断和科研经费申请被拒等(Hammarfelt, 2017; Kwiek & Roszka, 2021a, 2021b; Shibayama & Baba, 2015; Sugimoto & Larivière, 2023; Tang & Horta, 2023)。在大多数 STEM 学科领域,女性进入的仍是男性主导的环境——“老男孩”关系网络(old boy networks)营造的“寒冬氛围”(Santos et al., 2020)。

尽管男性和女性科学家退出学术界的比例均较高——在所有发表论文的科学家中,约三分之一的科研人员在从业前五年退出,近半数在十年内退出(后文将展示 2000 年队列的实证数据)——但学界普遍认为女性科学家流失率高于男性(Preston, 2004; Kaminski & Geisler, 2012)。“管道泄漏模型”(leaky pipeline model)与“寒冷气候假说”(chilly climate hypotheses)解释了 STEM 学科的性别差异:由于女性面临的系统性障碍,学术职业管道的每个阶段都存在女性人才流失(Blickenstaff, 2005; Goulden et al., 2011; Shaw & Stanton, 2012),而充满敌意或不受欢迎的工作环境会阻碍女性追求自己的职业生涯(Cornelius et al., 1988; Spoon et al., 2023)。

管道泄漏模型揭示学术职业发展的二元路径特征:科研人员或通过逐级跨越学术生涯阶段持续发展,或因结构性障碍而完全退出学术领域。STEM 领域女性科研者面临的“寒冷气候效应”具有多维度生成机制——其本质既是特定学科女性占比偏低导致的关系人口学困境,更与院系层级的性别结构失衡(女性教师比例)、制度程序公平性缺失及系统性性别偏见密切相关(Maranto & Griffin, 2011, pp. 143—146)。值得注意的是,管道泄漏模型的政策价值已在实践中得到验证,为促进欧美校园的性别平等提供了理论支撑。

有关科学界女性的文献主要呈现两种视角(Branch, 2016; Fox, 2020):科学“管道”视角关注的是个人,而科学“路径”视角关注的是个人所处的组织结构。前者以线性发展模型为基础,将学术生涯解构为可识别的连续阶段,其分析重心在于个体如何突破各阶段的结构瓶颈;后者则将学术发展视为嵌入制度网络的非连续进程,着重揭示院系治理模式、资源分配机制及权力关系等组织结构要素如何动

态塑造科研人员的发展轨迹(Fox & Kline, 2016, p. 63)。

“管道”隐喻突出了教育阶段之间以及教育与职业成果之间的直接联系(Xie & Shauman, 2003)。相关文献将研究焦点置于女性科研人员的个体特征差异(相较于男性群体),认为通过修正所谓的“能力缺陷”或“职业承诺不足”即可降低其学术流失风险。这种分析视角将结构性制度环境置于次要地位,忽视了女性作为学生、博士生至科学家的全程发展均受制于系统性约束的事实。“管道”隐喻的主要观点为“将女性保留在管道内”(Fox & Kline, 2016, p. 57),其隐含假设包含双重方面的简化:既预设学术职业为被动流经标准化阶段的机械过程,又将解决方案窄化为早期教育阶段的女性基数扩张(如增加更多的女性本科生、博士生及青年学者的性别比例)。这种范式本质上默认学术系统具备程序公平性,却回避了对“何种群体更易流失”及“流失机制根源”的深度追问(Branch, 2016, p. 7)。

“管道”隐喻存在如下问题:其一,未能解构科学共同体隐含的准入资格标准——即关于谁天生属于科学、谁天生被排除在科学之外、谁是刻板印象中理想的科学家(如将数学、计算机等领域与男性特质刻板关联);其二,它忽略了性别障碍,无视针对女性的“持续排斥信息”(Branch & Alegria, 2016, p. 8)。“管道”研究视角将科学界女性代表性不足归因为供给端基数不足与阶段性人员流失比例过高的简单叠加,其政策方案呈现出显著的技术治理倾向——既包括通过优化培养环节“堵塞管道泄漏”(Xie & Shauman, 2003, p. 7),也涉及早期职业阶段实施女性规模扩张策略(Fox & Kline, 2016, p. 63)。与此相反,“路径”隐喻强调既不直接也不简单的科学进步过程,将分析焦点转向科学家所处的组织文化及其工作环境。对于科学进步而言,重要的不仅仅是个人特质,而是女性日常从事科学工作的实际组织环境。组织文化涉及组织内的共同价值观、信仰和行为;组织的主导性文化定义了组织中更受重视(和不受重视)的东西,并提倡其中的“经营之道”。“管道”和“路径”两种视角衍生出本质迥异的解决方案:“管道”视角聚焦个体适应性提升,通过规训策略使女性科研者内化传统男性主导的科研实践范式(如推崇高强度工作伦理与去家庭化的职业承诺);“路径”视角则主张系统性制度重构,推动学术机构对性别、职级和奖励方面的不平等做出更积极的回应(Branch, 2016; Fox & Kline, 2016)。

最后,需明确区分作为性别分析工具的“寒冷气候”假说与学术组织场域中更具结构性的“院系气候”概念。后者源于科学实践的社会建构本质——科学家虽具个体能动性,但其职业发展始终嵌入制度化的组织条件网络(Fox, 2010)。科研工作本质上是集体协作与资源依赖的社会过程(Fox & Mohapatra, 2007)。因此,解析女性学术地位必须穿透个体层面,系统考察其所在机构的组织特征。正如Fox(2010, p. 1000)所述:“科学工作从根本上说是社会性和组织性的”。这些组织因素包括讨论研究的频率、关于研究的互动和交流、可用于研究的空间、在任职机构获得的认可和奖励,以及工作与家庭之间的干扰等,这些都影响着科学工作的参与和表现。科研机构的“个性”“特征”或“气候”,以及科研人员对所在机构文化的感知影响着科学活动。“院系气候”包括了科研人员对组织环境中价值观和实践的看法,可以通过调查来衡量(见Fox & Mohapatra, 2007, p. 141)。

尽管在理想情况下,学术生涯研究可以将个人层面和组织层面的分析结合起来,定量研究成果可以与定性研究成果一起进行检验,但受限于变量范围,本文的研究重点是描述性的,采用的是定量的、个人层面的数据。本文无法获得与机构及其细分单元相关的变量、除文献计量信息之外的个人特征变量,也无法了解个人对其所在部门工作方式的看法(气候、文化或基层研究组织的“工作方式”)。这些方面的分析需要对学术界进行一次全球性调查来解决。但是,本文能够做的是对大样本科学家进行二十多年的跟踪调查,并纵向分析其发表生涯中的性别差异。

二、研究设计:全球数据集与科学家研究的大数据方法

机构和国家层面的科学人才流失现象均可被量化分析,这一研究范式的革新得益于全球学术数据库的开放获取进程加速。越来越多的人可以使用数字数据库来获取有关科学家、他们的研究成果以及他们在全局学术对话中的引文影响的全面信息(Kashyap et al., 2023; Wang & Barabási, 2021),这为研

究全球和纵向的学术职业生涯提供了可能。新数字数据集的出现、巨大计算能力的获取以及社会科学研究中更普遍地转向结构化大数据, 导致近期有关学术职业各个方面的研究激增, 其中令人印象深刻的研究是从不同角度关注科学领域男女之间的差异 (King et al., 2017; Nielsen & Andersen, 2021; Sugimoto & Larivière, 2023)。大型数据集提供了独特的能力来检验有关科学和科学家的传统信念和概念框架 (Liu et al., 2023), 整个科学事业的追踪数字数据可用于多细节、大规模地捕捉科学的内部运作机理 (Wang & Barabási, 2021, p. 1)。现在, 虽然唾手可得的海量数据集也存在新的局限性 (Liu et al., 2023; Sugimoto & Larivière, 2023), 但系统探索成千上万名科学家的学术发表历史及其职业发展细节已经成为可能。

(一) 科学人才流失成为学术主题

然而, 无论是在全球范围内还是从多国视角来看, 都还没有对“离开科学界”这一主题进行全面研究。传统研究文献对这一问题的探讨主要通过小规模案例研究 (大多是基于调查和访谈的研究), 或者通过高校教师群体的多年期观察研究 (Rosser, 2004; Xu, 2008; Zhou & Volkwein, 2004)。如 White-Lewis 等 (2023) 对 2015 年至 2019 年期间 2 289 名美国高校离职教师进行了实证调查。总体而言, 这些研究普遍得出, 每个职业年龄段上的女性离开科学界的比例都高于男性, 并且相较于工作与生活的平衡, 女性离职决策更易受工作场所氛围的影响 (Spoon et al., 2023)。我们的研究结果也基本证实了最近的一项美国案例研究 (Spoon et al., 2023)。该研究得出的结论是女性在各个职业阶段离开学术界的比例总体上高于男性, 声望较低的机构尤其如此。

之前不少研究曾通过“教师离职意向”(faculty departure intentions) (Zhou & Volkwein, 2004)、“离职意向”(intentions to leave) (Rosser, 2004)、“教师流动”(faculty turnover) (Ehrenberg et al., 1991) 和“教师流失意向”(faculty turnover intentions) (Smart, 1990; Xu, 2008) 等概念对科学人才流失进行过研究。但大多数研究都集中在单一机构, 其地理范围也仅限于美国。Milojevic et al. (2018) 的研究是例外, 他们利用天文学、生态学和机器人技术领域的主要期刊的文献计量数据, 进行了全球范围的研究。以往研究表明, 女性更强烈的离职意愿与学术文化高度相关, 因为就男性和女性在家庭责任方面的角色而言, 学术文化为女性提供的晋升机会相对较少, 研究支持也很有限 (Xu, 2008)。然而, 学科差异也很重要, 因为不同学科的学者表现出不同的态度和行为模式, 学术环境内外提供的机会也各不相同 (Zhou & Volkwein, 2004)。

有研究指出, 学者对工作生活的看法直接影响到他们的满意度, 进而影响到他们的离职意向 (Rosser, 2004)。学者留在或者离开目前的机构受到多种内部力量的推动, 可分为三大类因素: 组织特征、个人特征和工作经历。这些因素会影响教职员工的工作满意度, 进而影响离职意愿。他们也会受到多种外部因素的推动, 离开他们的机构。内部因素包括个人和家庭特征 (如性别和家庭/婚姻状况)、组织特征 (如财力或工会) 和工作经历 (如工作量、生产率和报酬)。外部因素包括: 外部就业市场、外在奖励、研究机会、教学机会和家庭因素等。内部因素直接影响工作满意度和对组织环境的看法, 从而影响离职意向; 外部因素也被证明会加强或削弱离职意向 (Zhou & Volkwein, 2004, p. 144—147)。

做出职业流动的决定需要比较在当前机构就职和在其他机构就职的金钱和非金钱条件的预期现值 (Ehrenberg et al., 1991)。从经济角度来看, 根据薪资结构, 可能会出现不同的去留决定: 在薪资差距较小的机构中, 生产力最高学者可能倾向于离开 (因为他们觉得薪酬过低), 而在薪资差距较大的机构中, 生产力较低学者可能倾向于离开 (因为他们觉得与同事相比薪酬过低) (Ehrenberg et al., 1991, pp. 107—108)。Smart (1990) 发现, 科研时间更长、科研产出效率更高的男性终身教职人员离职意愿更强, 而薪资满意度仅对非终身教职员工有影响。

研究高校教师离职缘由的概念框架包括“推力”和“拉力”(当前学术工作场所的特征和吸引教师离开高教界的外部机构特征)。然而, 如果推力因素微乎其微, 那么拉力因素在离职决定中的权重就会

降低(White-Lewis et al., 2023)。离开科学界的原因包括工作与生活的平衡问题(Rosser, 2004; Smart, 1990)、工作保障低、工资低(Zhou & Volkwein, 2004)、同事和工作量问题(Wohrer, 2014)以及工作场所的各种歧视(Preston, 2004; Smart, 1990)和恶劣的工作氛围(Cornelius et al., 1988; Spoon et al., 2023)。

“离开机构”和“离开科学界”“离开意向”和“实际做出离开决定”(White-Lewis et al., 2023)之间存在重要的概念差异。本文的“离开科学界”是通过“不再在学术期刊上发表文章”来考察的。本文在生存分析中关注的重点是学者多年来的发表情况,直到某个时间点最终出现“不再发表”的情况,这已经超越了机构和部门的范畴,上升到了更广泛的“有学术发表的学者”和“无学术发表的学者”的层面。

(二) 数据与方法: 定义学术发表事件和生存分析

科学生涯可以概念化为一系列学术发表事件,从首次发表事件到后续发表事件,再到最近的发表事件,很多情况下,最近最新的发表事件也意味着科学家停止发表的时间事件。本研究中,2000年首次发表论文的科学家组成了2000年队列,在这个队列中,男性和女性科学家以不同的强度逐年逐渐减少。“离开科学界”被概念化为一个事件,并在所谓的生存分析方法中进行分析(Allison, 2014; Mills, 2011)。虽然“离开科学界”这一主题已被广泛探索(Geuna & Shibayama, 2015; Preston, 2004),但以停止出版为标志的“离开科学”这一事件尚未使用生存分析方法进行广泛研究。据我们所知,此前也未从大规模多国定量角度进行过研究。

在生存分析中,要探讨的问题与事件发生的时间和跨度有关(Mills, 2011)。本研究关注的重要事件是学者发表最后一篇论文的事件,标志着学者留在科学界的最后一年。离开科学界之前的时间跨度被称为生存时间(survival time)。生存经典的统计技术是用Kaplan-Meier估计生存时间(Mills, 2011)。Kaplan-Meier估计量的曲线图是一系列不同高度不等的水平阶梯。当研究中的一些受试者在研究结束前没有经历该事件时(就像本研究中,一些科学家在研究期间继续发表文章,根据“离开科学界”的定义,他们并未经历离开科学界的事件),这些受试者被称为右删失观测值。对于右删失观测数据,我们只能获得部分信息:受试者可能在研究结束前的最后一年或者研究结束后的任何时间经历该事件,但具体的事件经历时间并不可知。

本研究中,最后一次发表文章的日期为2019年或之后的作者被标记为删失观察值(即在2020年或之后的某个时间离开科学界)。若要将作者归类为离开科学界或留在科学界,其最后发表文章的日期必须是2018年(标记为2019年离开科学界)或更早。无删失案例是指同时知道学术职业生涯起始年份(2000年队列为2000年,2010年队列为2010年)和留在科学界的终止年份的观测数据,终止年份必须是2019年或更早(由最后一次发表文章的日期和第二年离开科学界的日期决定)。

对于每一年,进入时间间隔的初始科学家数量包括在这一时间区间内离开科学界的科学家和在下一时间区间内留在科学界的科学家。通过将给定时间间隔之前的所有时间区间内的所有生存概率相乘,可以计算出这一时间间隔内的总生存概率(Mills, 2011)。

相较于既往研究,本研究的创新性体现在地理尺度的突破(整合38个OECD国家的数据),摆脱了依赖问卷调查与深度访谈、聚焦单一国家机构的个案研究范式,实现了向多国别、学科导向型研究的跨越式转变。在研究设计层面,本研究通过持续追踪大量非重叠科学家群体长达二十年,系统揭示了科研人员跨学科流动与性别差异的历时性演变规律。特别值得指出的是,本研究采用Scopus数据库提供的结构化、可靠且经过严格筛选的大数据资源,研究团队摒弃传统的横截面研究(单一时点观测)或重复横截面研究法(多时点独立观测),转而构建了2000—2022年间连续二十年的科学家年度追踪观测数据。研究过程中始终确保不同国家采用相同的研究变量和稳定的分类方式,整合了多维度的个体微观数据,并将时间变量作为核心分析维度贯穿研究全程。

(三) 数据集和数据流: 从出版物元数据中了解科学家个体信息

本研究基于Scopus数据库中2000年至2010年间首次发表学术成果的科研人员出版物文献计量

数据,系统梳理了该时间段内全球学术生产活动的量化特征。研究样本覆盖 2,127,803 名活跃学者,其学术产出与引用关系通过标准化的 Scopus 作者标识符(Author ID)实现精准关联,为跨国、跨性别的微观学术计量分析提供了可靠数据基础(Baas et al., 2020)。研究框架如图 1 所示,样本群体按学术生涯起点划分为两组:左侧为 2000 年开启研究生涯的学者群体,右侧为 2010 年进入学术领域的新生力量。为确保数据质量,本研究对样本施加了以下筛选条件:纳入研究范围的科学家需满足其出版作品集至少包含两篇论文(或会议论文),所属国家限定为经济合作与发展组织(OECD)成员国,性别基于可获得数据明确标注为男性或女性,学科领域聚焦于科学、技术、工程、数学与医学(STEMM)。这种双时点对照设计不仅捕捉了学术生命周期的动态变化,更为分析十年间全球知识生产格局的演变提供了独特视角。

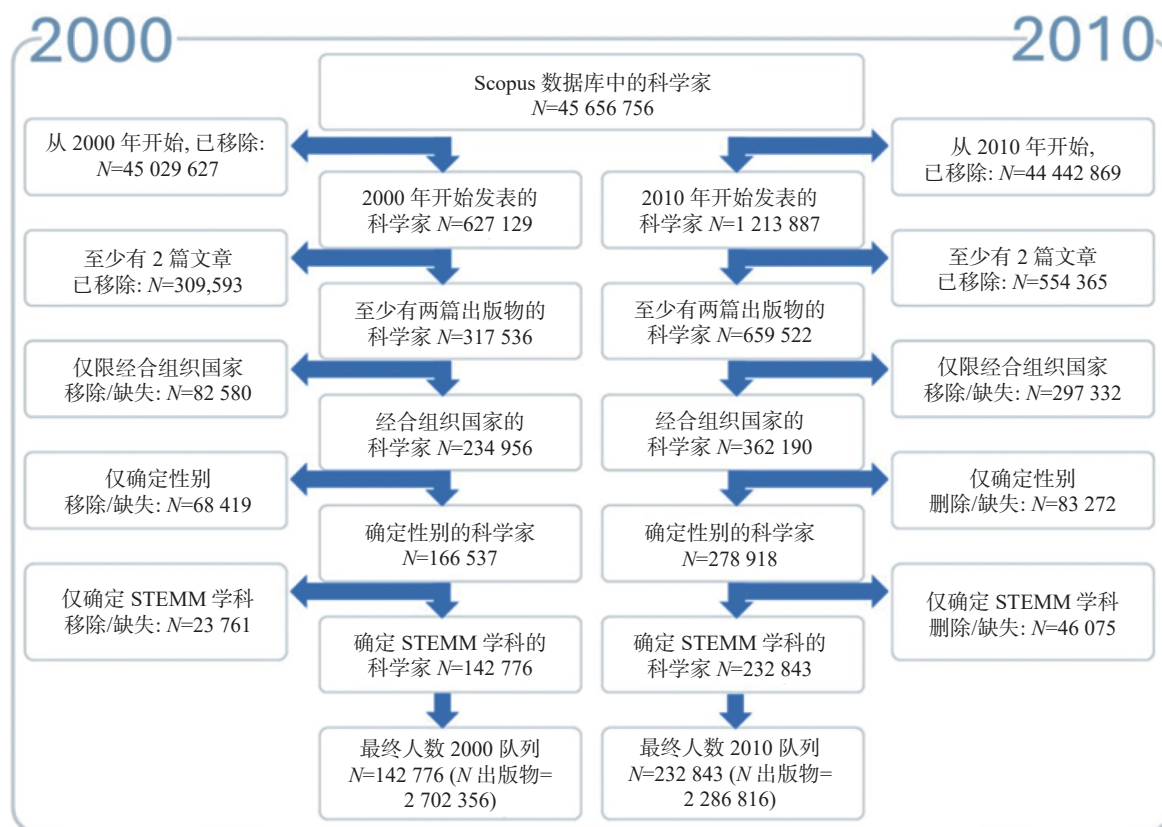


图 1 2000 年(左)和 2010 年(右)科学家群体的筛选过程

本研究依据 Scopus 数据库的期刊分类系统(ASJC)定义的 16 个 STEMM 学科进行分类,具体包括:农业与生物科学(AGRI)、生物化学与分子生物学(BIO)、化学工程(CHEMENG)、化学(CHEM)、计算机科学(COMP)、地球与行星科学(EARTH)、能源(ENER)、工程(ENG)、环境科学(ENVIR)、免疫学与微生物学(IMMU)、材料科学(MATER)、数学(MATH)、医学(MED)、神经科学(NEURO)、药理学与毒理学(PHARM)、物理学与天文学(PHYS)。研究对样本中的科学家进行了性别、学科、国家以及退出学术出版活动(而非退出科学部门或高等教育就业)年份的赋值。

其一,性别的确定。性别确定算法的有效性对于科学界的性别比较研究至关重要。因为同行评审的出版物通常不会披露作者的性别(Science-Metrix, 2018),在大规模多国研究中,利用作者姓名信息推断性别是唯一可行的手段。然而,这种方法存在局限性:一是部分论文仅提供作者首字母缩写;二是可获取姓名的论文比例随时间、领域和国家不同而变化;三是并非所有名字都具有性别区分特征。本研

究选择 Scopus 数据库而非 Web of Science, 是因为后者在 2008 年之前无法生成国家层面的性别指标 (Science-Metrix, 2018), 而本研究的时间跨度为 2000 年至 2022 年。

在之前国家层面的较小规模研究中(N=25,463 名科学家, Kwiek & Roszka, 2021a; Kwiek & Roszka, 2021b; Kwiek & Roszka, 2022), 我们使用了行政和传记数据集(来自政府部门国家科学家登记处的数据)提供的性别信息;而在基于调查的研究中,我们使用了科学家自我申报的性别信息(Kwiek, 2016; Kwiek, 2018)。但本研究涉及的 OECD 国家没有个人微观层面的行政数据。因此,为了确定科学家的性别,我们采用了爱思唯尔最近两份关于科学界女性的报告(Elsevier, 2018; Elsevier, 2020)所采用的方法。作者的性别推断数据是根据多年合作协议,由国际研究中心(ICSR)实验室平台提供。

其二,学科的确。本研究将作者学术生涯中所有引用论文所在期刊的主导学科确定为作者的学科领域。具体方法如下:首先从作者各类出版物组合中提取所有引用文献(共计 2,092,766,869 篇),每篇引用论文均标注了基于 4 位数 ASJC 代码的学科分类(见图 1)。根据 ASJC 代码的前 2 位数值来确定论文所属的学科,并剔除重复值以确保唯一性。随后统计每位作者引用各学科论文的频次,将其引用次数最多的学科(众数)作为主导学科。若作者未被分配学科或存在两个及以上主导学科,则将其从样本中剔除(采纳样本 N=32,794,309,剔除样本 N=12,862,447)。本研究未采用随机分配学科的方法,而是进一步聚焦于 STEMM 领域作者,最终形成包含 29,927,584 名作者的分析样本(采纳样本 N=29,927,584,剔除样本 N=2,866,725)。这种基于引用行为的学科分类方法避免了单纯依赖出版期刊的局限性,能够更准确地反映学者的研究专长分布。通过标准化的 ASJC 代码转换规则,确保了学科分类的全球一致性,同时剔除了多学科交叉导致的模糊归属问题,为后续的性别与学科交叉分析奠定了坚实的数据基础。

其三,国家的确定。本研究采用作者出版物中标注的主要国家作为其国家分类依据。具体操作如下:从作者的各类出版物中提取其注明的所有所属国家,并统计每位作者标注每个国家的频次。在确定主要国家时,选择作者发表各类出版物数量最多的国家(众数)。与学科分类类似,未能被分配国家或存在两个及以上主要国家的作者被剔除(采纳样本 N=39,405,552,剔除样本 N=6,251,204)。随后,将样本进一步限定为 38 个 OECD 国家的作者(采纳样本 N=23,619,928,剔除样本 N=15,785,624)。

其四,作者退出学术出版活动的年份的确定。研究基于每位作者在 Scopus 数据库中的唯一标识符及其最后一次发表任何类型论文的年份。将最后一次发表论文年份的次年定义为退出出版年(即最后一次发表论文年份+1)。若作者最后一次发表论文的年份为 2018 年或更早,则 2019 年为研究的最后一年,其退出出版年份在 2019 年之后(2020 年、2021 年、2022 年)的作者被归类为删失观测值。这种基于出版物记录的国家退出年份分配方法,确保了数据的客观性与一致性,同时通过剔除多国归属和多学科交叉的模糊样本,为后续分析提供了清晰且具有代表性的研究群体。

三、研究结果及分析

首先,2000 年队列中的 142,776 名 STEMM 科学家的论文发表情况从 2000 年追踪至其停止发表论文(或至 2022 年)。图 2 展示了所有 STEMM 学科的 Kaplan-Meier 生存曲线图其中 y 轴表示持续从事科学研究的科学家比例(即继续发表论文的概率),x 轴表示自 2000 年以来发表论文的年数。为确保数据完整性,2020—2022 年的右删失样本以小交叉标记表示。

男性和女性的生存曲线在学术生涯的前四年(通常是攻读博士学位的阶段)均出现显著下降,表明这一时期是科学家流失的高峰期。从第四年开始,性别差异逐渐显现:男性科学家每年持续发表论文的比例开始超过女性,这一趋势与以往研究结果相符。具体来看,2000 年进入学术领域的 STEMM 科学家群体中,约三分之一在 5 年内离开科学界,约一半在 10 年内离开,约三分之二在 19 年研究期结束时离开。整个研究期间,男性离开科学界的概率始终低于女性。具体数据表明(见表 1),女性在 5 年

和 10 年后离开科学界的概率分别比男性高出 12.54% 和 11.52%，而在研究期结束时，这一差距为 6.33%。对于未删失观测值，女性在 19 年后继续发表论文的 Kaplan-Meier 估计概率为 29.4%，而男性则高达 33.6%。然而，这些总体趋势掩盖了不同学科间的显著差异（见图 3）。尽管 STEMM 领域整体呈现人才流失现象，但各学科的流失模式存在重要差异。

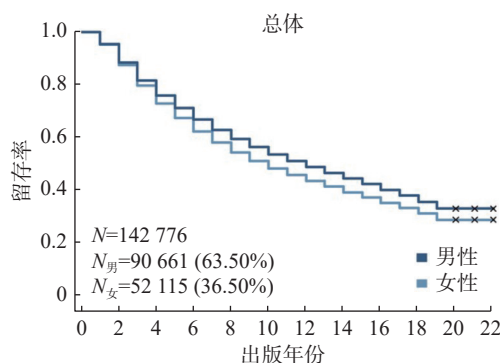


图 2 2000 年科学家队列（n=142,776）按性别列的 Kaplan-Meier 生存曲线图

表 1 按性别列的 2000 年科学家群体的 Kaplan-Meier 估计值（所有学科合计）

时间 (年)	女性（2000 年队列）			男性（2000 年队列）			总计（2000 年队列）		
	N	N（离开 科学界）	KM 概率（停留） 及 95% CI 和 SE	N	N（离开 科学界）	KM 概率（停留） 及 95% CI 和 SE	N	N（离开 科学界）	KM 概率（停留） 及 95% CI 和 SE
1	52,115	2,530	0.951 (0.950-0.953) ¹	90,661	4,151	0.954 (0.953-0.956) ¹	142,776	6,681	0.953 (0.951-0.954) ¹
2	49,585	3,985	0.875 (0.872-0.878) ¹	86,510	6,302	0.885 (0.883-0.887) ¹	136,095	10,287	0.880 (0.877-0.882) ¹
3	45,600	3,948	0.799 (0.796-0.803) ²	80,208	6,114	0.817 (0.815-0.820) ¹	125,808	10,062	0.811 (0.809-0.812) ¹
4	41,652	3,553	0.731 (0.727-0.735) ²	74,094	5,062	0.761 (0.759-0.764) ¹	115,746	8,615	0.749 (0.746-0.751) ¹
5	38,099	2,838	0.677 (0.673-0.681) ²	69,032	4,356	0.713 (0.710-0.716) ²	107,131	7,194	0.695 (0.691-0.698) ²
6	35,261	2,602	0.627 (0.623-0.631) ²	64,676	3,934	0.670 (0.667-0.673) ²	99,937	6,536	0.655 (0.651-0.658) ²
7	32,659	2,183	0.585 (0.581-0.589) ²	60,742	3,458	0.632 (0.629-0.635) ²	93,401	5,641	0.613 (0.610-0.616) ²
8	30,476	1,961	0.547 (0.543-0.551) ²	57,284	3,110	0.598 (0.594-0.601) ²	87,760	5,071	0.577 (0.574-0.580) ²
9	28,515	1,665	0.515 (0.511-0.520) ²	54,174	2,774	0.567 (0.564-0.570) ²	82,689	4,439	0.548 (0.545-0.551) ²
10	26,850	1,472	0.487 (0.483-0.491) ²	51,400	2,465	0.540 (0.537-0.543) ²	78,250	3,937	0.517 (0.514-0.520) ²
11	25,378	1,264	0.463 (0.458-0.467) ²	48,935	2,225	0.515 (0.512-0.518) ²	74,313	3,489	0.492 (0.489-0.495) ²
12	24,114	1,158	0.440 (0.436-0.445) ²	46,710	2,055	0.493 (0.489-0.496) ²	70,824	3,213	0.466 (0.463-0.469) ²
13	22,956	1,151	0.418 (0.414-0.423) ²	44,655	2,032	0.470 (0.467-0.473) ²	67,611	3,183	0.444 (0.441-0.447) ²
14	21,805	1,089	0.398 (0.393-0.402) ²	42,623	1,889	0.449 (0.446-0.453) ²	64,428	2,978	0.426 (0.423-0.429) ²
15	20,716	1,048	0.377 (0.373-0.382) ²	40,734	1,884	0.429 (0.425-0.432) ²	61,450	2,932	0.405 (0.402-0.408) ²
16	19,668	1,033	0.358 (0.353-0.362) ²	38,850	1,959	0.407 (0.404-0.410) ²	58,518	2,992	0.384 (0.381-0.387) ²
17	18,635	1,002	0.338 (0.334-0.342) ²	36,891	2,020	0.385 (0.381-0.388) ²	55,526	3,022	0.363 (0.360-0.366) ²
18	17,633	1,064	0.318 (0.314-0.322) ²	34,871	2,070	0.362 (0.359-0.365) ²	52,504	3,134	0.342 (0.339-0.345) ²
19	16,569	1,228	0.294 (0.290-0.298) ²	32,801	2,350	0.336 (0.333-0.339) ²	49,370	3,578	0.315 (0.312-0.318) ²

注：(1) 标准误差为 0.001，(2) 标准误差为 0.002。表中包括男女总人数、时间（以年为单位）、离开科学界的科学家观察数量和留在科学界的生存概率，在 95% 的置信区间内（下表同）。

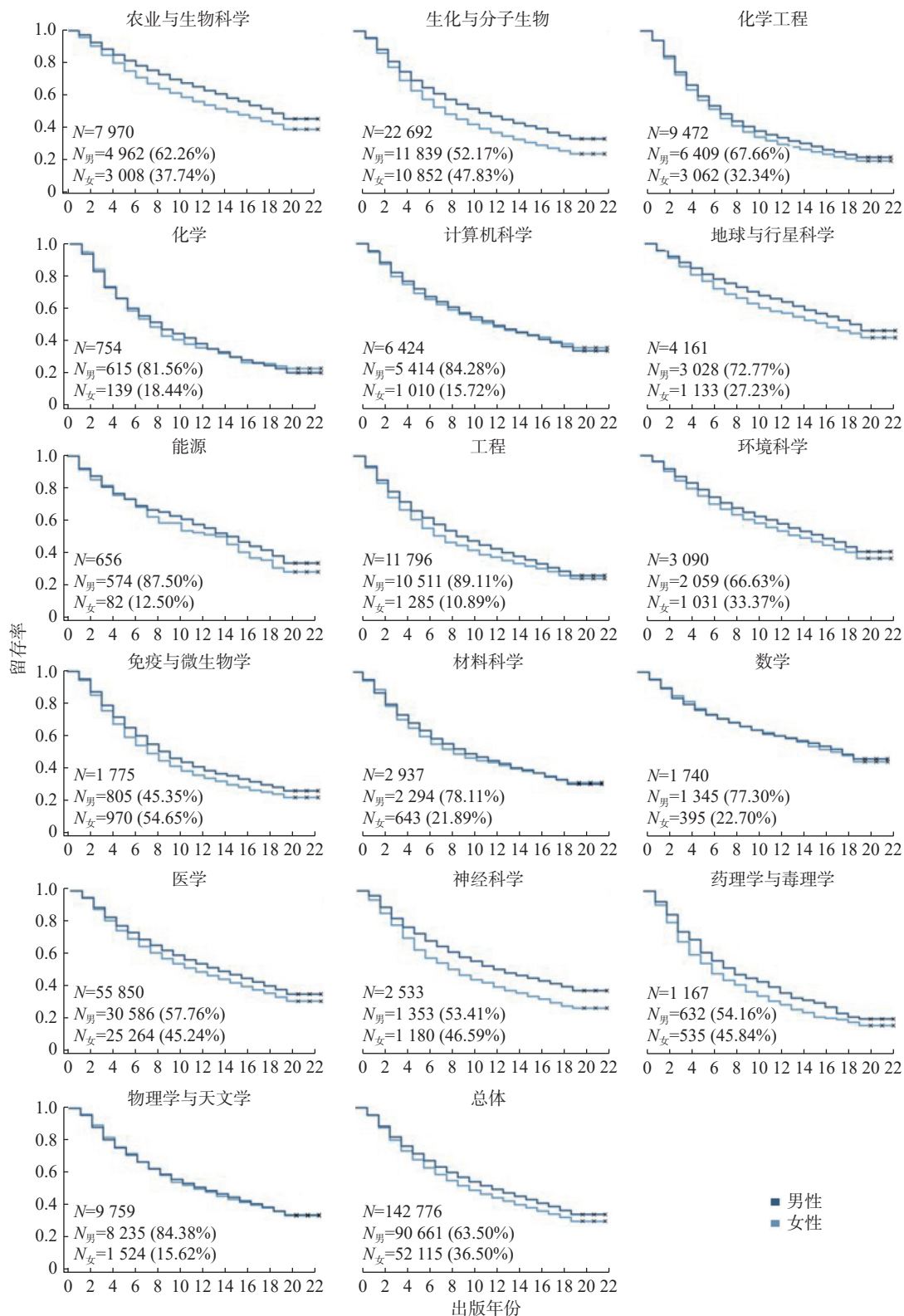


图 3 2000 年科学家队列 (N=142, 776) 按学科和性别分列的 Kaplan-Meier 生存曲线图

2010 年队列的男性和女性科学家的 Kaplan-Meier 生存曲线明显与 2000 年队列不同。在所有学科中, 无论是否为数学密集型学科(如计算机科学、物理学和工程学), 男女科学家的生存曲线几乎完全

重合。具体到 2000 年队列的详细 Kaplan-Meier 估计(见表 1), 女性科学家在学术生涯第一年留在科学界的概率为 95.1%(95% 置信区间: 95.0%-95.3%), 第二年降至 87.5%, 第四年进一步下降至 73.1%, 意味着四年内累计离开学术界的概率达到 26.9%。五年后, 女性留在科学界的概率为 67.7%, 十年后为 48.7%, 至研究期结束(19 年后)仅为 29.4%。相比之下, 男性科学家在相同时间点的相应概率分别为 71.3%、54.0% 和 33.6%, 均显著高于女性。

逐年数据显示, 科学家离开学术界的高峰期通常出现在职业生涯的前 6 至 8 年。同时, 科学家首次发表后第二个十年的逐年留任概率通常高于第一个十年。女性在第 10 年和男性在第 12 年的 Kaplan-Meier 留任概率均低于 50%。女性在科学界生存 15 年及以上的概率为 37.7%, 而男性为 42.9%。

对于 2010 年队列(见表 2), 进入科学界 9 年后, 女性留下的概率为 0.414(队列中 41.4% 的女性持续有学术发表); 对于男性来说, 留下来的概率仅略高一点, 为 0.424(42.4%), 与 2000 年队列的结果相比, 这种差异非常小, 而 2000 年队列的结果存在显著的性别差异。在包括数学密集型学科(计算机、工程学、数学和物理)在内的八个学科中, 统计检验表明, 男性和女性的生存曲线没有统计学上的显著差异。但在所有学科综合分析中, 以及在女性约占 50% 的医学、生物和农业三个学科中, 差异在统计上是显著的(通过六种方法来验证结论的稳健性, 结果未在此报告)。

表 2 按性别分列的 2010 年科学家群体的 Kaplan-Meier 估计值 (所有学科合计)

时间 (年)	女性 (2010 年队列)			男性 (2010 年队列)		
	N	N (离开科学界)	KM 概率 (停留) 及 95% CI 和 SE	N	N (离开科学界)	KM 概率 (停留) 及 95% CI 和 SE
1	97,145	5,030	0.948 (0.947-0.950) ¹	135,698	7,375	0.946 (0.944-0.947) ¹
2	92,115	8,686	0.859 (0.857-0.861) ¹	128,323	12,183	0.856 (0.854-0.858) ¹
3	83,429	8,090	0.776 (0.773-0.778) ¹	116,140	11,164	0.774 (0.771-0.776) ¹
4	75,339	7,369	0.700 (0.697-0.703) ¹	104,976	9,869	0.701 (0.698-0.703) ¹
5	67,970	6,470	0.633 (0.630-0.636) ²	95,107	8,552	0.638 (0.635-0.640) ¹
6	61,500	5,904	0.572 (0.569-0.575) ²	86,555	7,851	0.580 (0.577-0.583) ¹
7	55,596	5,499	0.516 (0.513-0.519) ²	78,704	7,322	0.526 (0.523-0.529) ¹
8	50,097	4,984	0.464 (0.461-0.468) ²	71,382	7,078	0.474 (0.471-0.477) ¹
9	45,113	4,929	0.414 (0.411-0.417) ²	64,304	6,745	0.424 (0.422-0.427) ¹

注: (1) 标准误差为 0.001, (2) 标准误差为 0.002。

在科学职业中, 我们特别关注男女生存曲线差异最大与最小的学科。医学(MED)和生物化学、遗传学和分子生物学(BIO)作为规模最大的两个学科, 同时也是女性科研人员占比最高的领域, 其科研人员的生存曲线呈现出显著的性别差异。以生物化学、遗传学和分子生物学(BIO)为例(见图 4 左图, 与右图形成对比), 该领域共涵盖 22,692 名科学家, 女性占比达 47.8%。在第 2-4 年期间, 生存曲线下降幅度最为显著。自第 3 年起, 生存曲线的男女差距逐渐扩大, 男性科学家留在科学界的比例变化更为平稳。2010 年科学家队列的 Kaplan-Meier 生存曲线图存在显著性别差异。具体来看, 女性科学家在进入科学界 9 年后继续发表论文的概率为 41.4%, 而男性科学家略高, 为 42.4%。与 2000 年队列相比, 这种性别差异有所缩小, 但在医学、生物和农业等学科中, 女性科学家的流失率依然较高。

在生物学领域, 女性科研人员在职业生涯各阶段都面临较高的流失风险。具体来看, 女性在进入该领域 5 年后离开科学界的概率高达 37.2%, 10 年后这一比例上升至 58.3%, 至研究期结束(19 年后)更是达到 76.6%。相比之下, 男性科学家在相同时间点的流失率分别为 30.8%、48.6% 和 67.3%。这意味着, 女性在 5 年和 10 年后离开科学界的概率分别比男性高出 20.8% 和 20.0%, 而在研究期结束时, 这一差距为 13.8%。物理学领域则呈现出不同的趋势, 是人员流失性别差异不明显的典型例子。在

物理学中,女性学者在 5 年后离开科学界的概率为 28.1%,与男性学者的 29.2% 相近;10 年后,女性的流失率为 47.9%,略高于男性的 46.9%;至研究期结束时,女性和男性科学家流失率几乎持平,分别为 66.9% 和 66.5%。

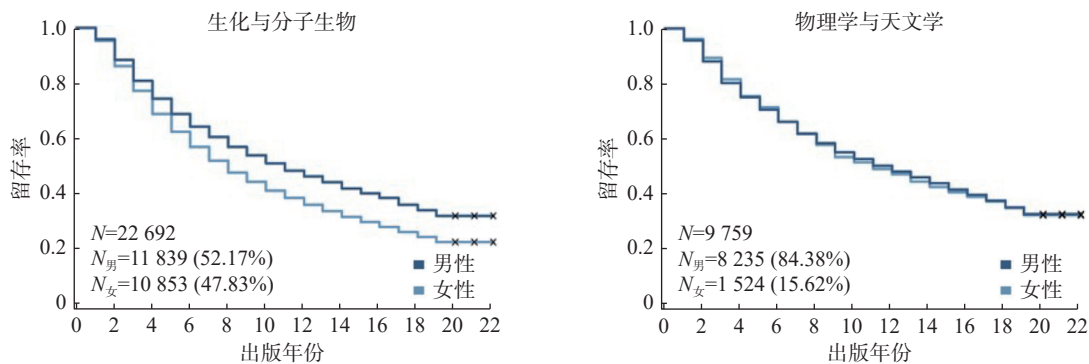


图 4 2000 年科学家队列中按性别分列的 Kaplan-Meier 生存曲线图
生物化学与分子生物学 (N=22,692, 左图) 与 物理与天文学 (N=9,759, 右图)

在数学、计算机科学和物理学这三个数学密集型学科中,尽管女性科研人员数量和比例相对较低,但不同性别科学家的 Kaplan-Meier 生存曲线图却呈现出高度一致性(见图 4 右侧的物理学曲线),两条曲线几乎完全重合。工程学领域的科研人员流失率有所不同,显示出一定的性别差异。基于 38 个 OECD 成员国的大量样本数据,研究进一步证实了在物理学、化学和天文学领域,人员流失不存在显著的性别差异。总体而言,在四个数学密集型学科中,性别差异并不具有统计学意义。然而,在女性科研人员占比较高的学科(如农业、生物和医学)以及所有学科的总体数据中,性别差异在统计上是显著的。

这种跨学科的对比如揭示了一个重要事实:性别差异在不同学科中的表现并不一致。在数学密集型学科中,尽管女性参与度较低,但一旦进入这些领域,她们的职业稳定性与男性并无显著差异。而在女性占比较高的学科中,性别差异则更为明显。这一发现提示我们在分析学术职业性别问题时,需要结合学科特性进行具体考察,而非简单地将性别差异归因于学科类型或数学密集度。

为了更全面地了解科学人才的流失和保留情况,我们将 2000 年队列和 2010 年队列中的男性和女性数据关联起来,以生成生存回归曲线(B)、风险率曲线(C)和核密度曲线(D),从不同角度分析科学界的人员流失与保留状况。生存分析用于研究相关事件的发生概率(本文指停止发表论文)。对于 2000 年和 2010 年队列中的所有科学家,研究具有相同的开始时间,而对于部分科学家,事件发生在研究开始后的一段时间内。风险率是指在研究期间任何给定时间点上发生事件的概率。我们计算了每位科学家每年停止发表论文的概率,并绘制了结果图。

生存回归函数是基于 Kaplan-Meier 估计值构建的模型,用于展示科学家在随后几年中仍在发表论文的比例变化。生存回归曲线实际上是 Kaplan-Meier 曲线的平滑版本。它从 y 轴上的 1 开始,因为所有科学家在时间零点都在发表论文,随后男性和女性科学家的比例逐渐下降至某一水平。与 Kaplan-Meier 生存曲线类似,2000 年和 2010 年队列中所有学科的生存回归曲线(图 5 和图 6 中的 B)显示,科学家在出版生涯早期的下降幅度较大,而在后期的下降幅度较为平缓。与 2010 年队列相比,2000 年队列中男性和女性的曲线差异逐渐扩大。具体而言,2000 年队列中,男性科学家在职业生涯后期的留存比例高于女性科学家,且男性和女性科学家在职业生涯后期的留存比例都高于职业早期。

对所有学科的概括性描述掩盖了一些更为细微的学科特异性描述。2000 年队列中的生物学和物理学领域的科学家在 Kaplan-Meier 曲线上呈现出截然不同的情况。这些发现在比较生存回归曲线时

得到了证实(图 7A), 两个学科的科学家人虽然在职业早期有较高的离开科学界的概率, 在职业后期有较低的离开科学界的概率, 但是在性别差异上存在显著不同。数据显示, 随着时间的推移, 女性生物学者从科学界离开的比例比男性更高; 相反, 在整个研究期间, 女性物理学者从科学界离开的比例几乎与男性相同。从跨组比较的角度来看, 各学科(以及所有学科加总后)的人才流失模式发生了根本性的转变。2000 年科学家群体中男女之间的差异非常明显, 但在 2010 年群体中这种差异几乎消失。

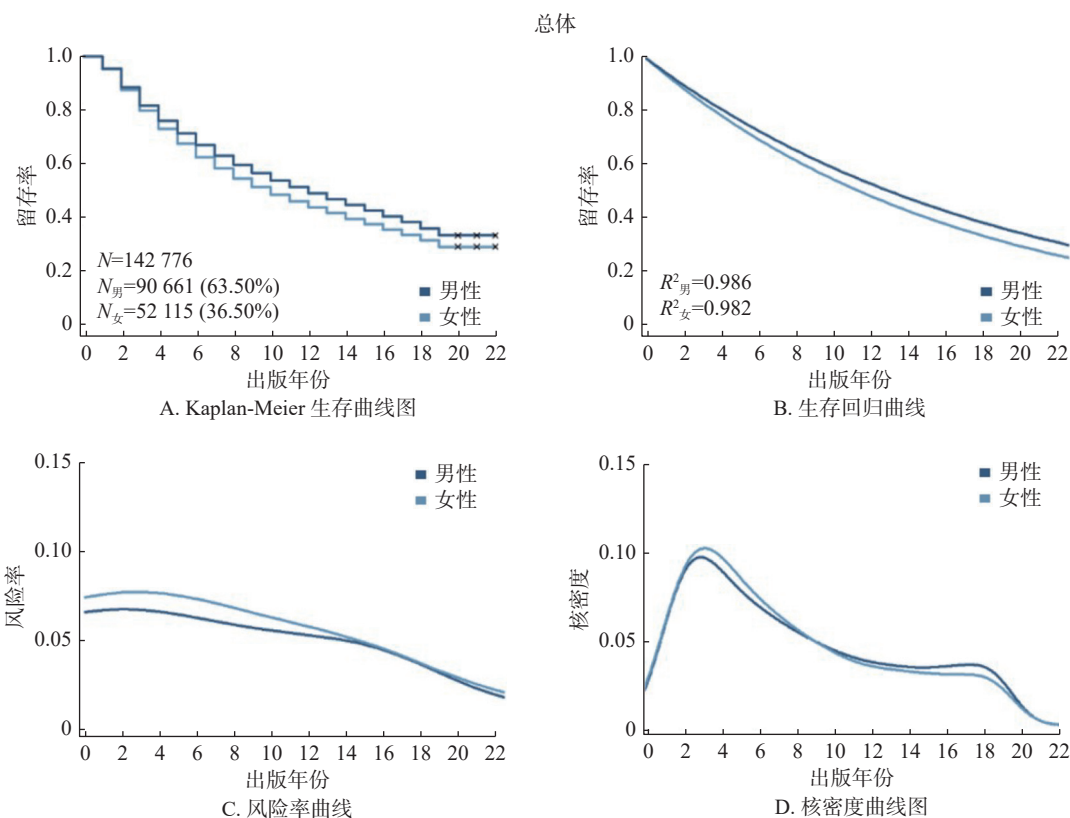


图 5 2000 年科学家队列 (N=142,776), 所有学科合并

A. Kaplan-Meier 生存曲线图 B. 生存回归曲线 (Kaplan-Meier 指数分布拟合) C. 风险率曲线 (样条平滑法, 平滑参数 10k) D. 核密度曲线图 (样条平滑法, 带宽 2, 基于高斯曲线的每点分量)

这一发现具有重要的潜在研究和政策意义。从事科学的时间(比如进入科学界十年前后的科学家群体的差异)是重要的考量因素。两组人群所有学科的 Kaplan-Meier 生存曲线呈现出截然不同的形式, 2010 年群体的流失率更高且下降幅度更大, 且无性别差异。生存回归曲线、风险率曲线和核密度曲线均证实了这一点, 表明在从事科学工作的前 4 年, 男性和女性科学家的流失率都极高。

2000 年和 2010 年科学家群体的生存分析显示了显著的差异。2000 年群体中, 生物科学领域的女性科学家流失率显著高于男性, 而物理科学领域的男性和女性科学家流失率几乎相同。2010 年群体中, 这种性别差异在生物科学领域有所缩小, 而在物理科学领域则依然保持较小差异。在生物科学领域, 2010 年队列的 Kaplan-Meier 生存曲线和生存回归曲线显示, 男性和女性科学家的差异比 2000 年队列小得多。风险率曲线和核密度曲线也显示出类似的差异, 尽管在前 10 年的强度不同。这表明, 随着时间的推移, 生物科学领域的性别差异在逐渐缩小。相比之下, 物理科学领域的性别差异一直较小。2010 年科学家队列的 Kaplan-Meier 生存曲线和生存回归曲线显示出较小的性别差异, 风险率曲线和核密度曲线也证实了这一点。核密度分布显示, 男性在第 2—4 年离开科学界的可能性高于女性, 表明男性更有可能在职业生涯早期离开。

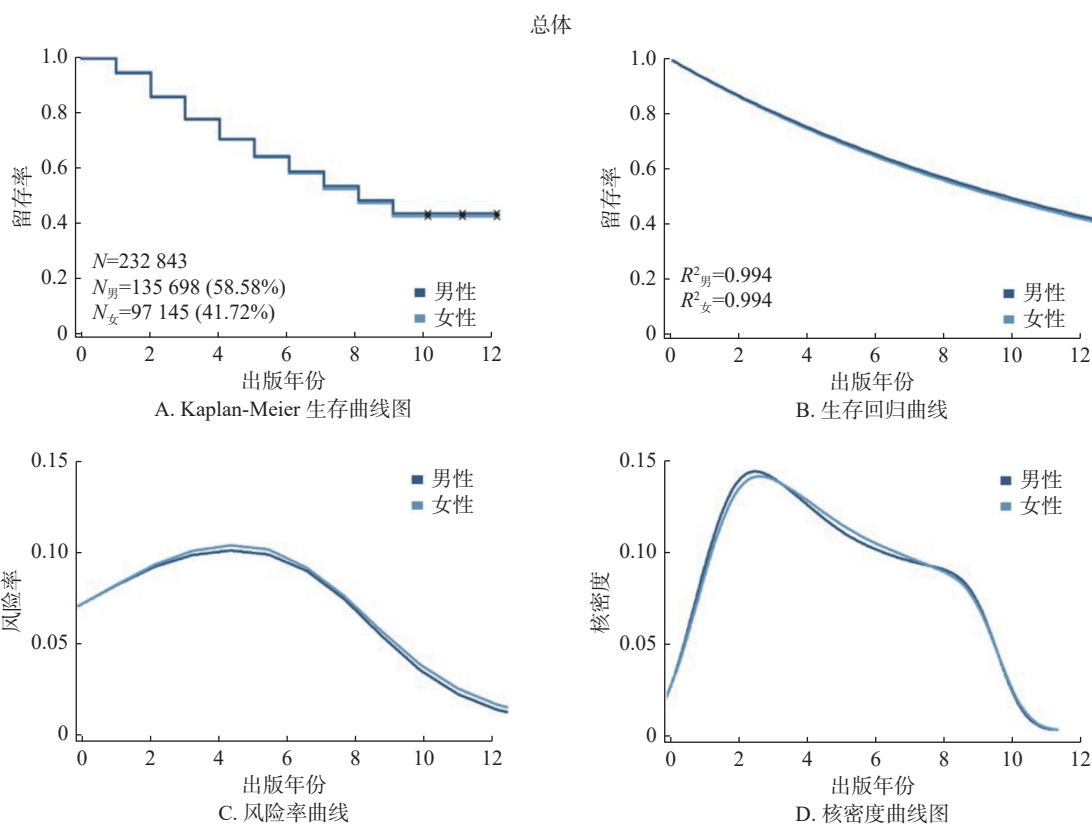


图 6 2010 年科学家队列 (N=232,843)，所有学科合并

A. Kaplan-Meier 生存曲线图 B. 生存回归曲线 (Kaplan-Meier 指数分布拟合) C. 风险率曲线 (样条平滑法, 平滑参数 10k) D. 核密度曲线图 (样条平滑法, 带宽 2, 基于高斯曲线的每点分量)

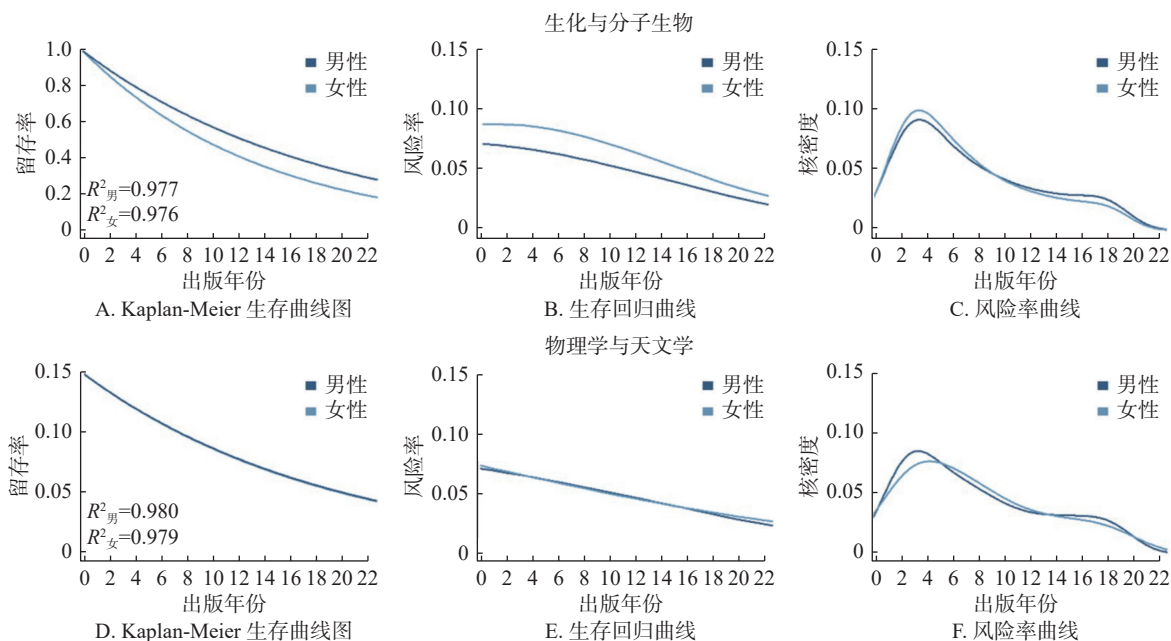


图 7 2000 年科学家队列中的生物 (N=22,692) 和物理 (N=9,750)

A. Kaplan-Meier 生存曲线图 B. 生存回归曲线 (Kaplan-Meier 指数分布拟合) C. 风险率曲线 (样条平滑法, 平滑参数 10k) D. 核密度曲线图 (样条平滑法, 带宽 2, 基于高斯曲线的每点分量)

四、结论与讨论

科学人才的流失和保留是一个长期的过程,要超越单一国家按学科和时间分析这一现象,需要大规模的多国纵向数据集。随着每年新一批科学家在新的条件下进入科学界,“离开科学界”正在经历重大转变(Milojevic et al., 2018)。

基于队列的纵向研究表明,在所有 STEMM 学科层面的综合变化之外,出现了广泛而细微的变化,这种变化在不同学科层面和不同时间范围内的强度各不相同。在“停止发表论文”的操作化定义下,“科学人才流失”对不同学科的男性和女性科学家有不同的含义,对不同时间进入科学界的科学家群体也有不同的含义。

2000 年科学家群体中约三分之一的人在 5 年后离开科学界,约一半在 10 年后离开,约三分之二在研究期结束时(19 年后)离开,男性离开科学界的比例始终较低,女性则较高。女性在 5 年和 10 年后离开科学界的可能性比男性高出约十分之一(分别为 12.54% 和 11.52%),在研究期结束时高出 6.33%。

与 2000 年进入科学界的科学家相比,2010 年进入科学界的科学家在流失方面的性别差异较小:随着更多的女性进入科学领域,科学人才流失的性别差异越来越小。在整合所有学科科学家数据的情况下,38 个 OECD 国家的 2000 年科学家队列的流失性别差异在 2010 年科学家队列中消失了。总体情况背后隐藏着这两批科学家细微的分类情况,不同学科有不同的持续进程。

生物化学、遗传学和分子生物学(BIO, 规模大, 女性占 47.83%)以及物理学和天文学(PHYS, 规模大, 女性仅占 15.62%),这两个对比鲜明的学科在 2000 年的 Kaplan-Meier 曲线中显示了截然不同的情况。在生物科学领域,女性科学家留任科学界的比例明显低于男性,而且随着时间的推移,差异越来越大;而在物理和天文学领域,二十年来男性和女性的留任比例几乎完全相同。与男性相比,二十年来,生物科学领域的女性从科学界消失的比例越来越大;相比之下,在整个研究期间,物理科学领域的女性从科学界消失的比例几乎与男性的相同。

对于 2010 年队列,所有学科中男性和女性的流失率差异明显较小,但不同学科之间仍然存在性别差异。在数学密集型学科,如数学、计算机科学和物理科学,性别差异几乎不存在。尽管这些领域中女性的比例很小,但她们在学术职业上的成功率与男性相同,表明她们在这些领域取得了极大的成功。这种现象可能与这些领域的性别化筛选机制有关,成功发表两篇研究文章的女性更有可能长期成为学术活跃科学家。在女性代表性非常低的学科中,女性通常扮演“象征性代表(token)”或院系标杆角色(Kanter, 1977),她们在科学体系中的地位与男性一样强大。这表明,尽管女性在这些领域中是明显的少数群体,但她们在学术职业上的表现与男性并无显著差异。这种现象可能与她们在这些男性主导领域中所面临的挑战和机遇有关,她们需要在高度竞争的环境中证明自己的能力,从而在学术界获得认可和地位。

以计算机为例,与其他 STEMM 学科相比,计算机专业中发表论文的女性人数随着时间的推移缓慢增加,而计算机专业中的女性在高级学术职级中却很少见,而这对她们的决策能力至关重要(Fox & Kline, 2016)。Branch 和 Alegria(2016)报告了女性在计算机领域面临的性别障碍和“持续的排斥性信息”。计算机、数学、物理和工程学的研究表明,这些领域的女性可能与 STEMM 领域的女性学者的平均水平不同,尤其是与女性人数众多的医学、生物和农业领域相比。这些学科的成功女性可能是“与众不同”“非典型”,而不是分布中心的普通女性(Branch & Alegria, 2016)。她们从副教授晋升为正教授的机会在很大程度上取决于被认为具有激励性/集体性的部门氛围(Fox & Xiao, 2013)。计算机科学中女性的学术职级(而不仅仅是存在)很重要,因为职级是大学中地位和决策权的归属(Fox & Kline, 2016)。

尽管传统上认为女性(Goulden et al., 2011; Preston, 2004; Shaw & Stanton, 2012)比男性更早、比例更高地离开科学界(这在我们对 2000 年队列的分析中得到了普遍证实),导致许多学科的流失率更高。但在相比之下,对于最近的群体,流失率的性别差异可能不再存在,尤其是对于女性人数和比例较低的

数学密集型学科。对于新一代科学家来说,科学界的流失率一直在上升,而且非常高(2010年群体中,58.6%的女性和57.6%的男性在9年内从科学界消失或停止发表文章),但这种性别差异似乎比传统假设的要小得多。

在过去的三十年中,女性参与科学研究的情况发生了显著变化,基于文献计量学的科学家纵向研究为全球跨性别和跨队列分析提供了新的机遇。在瞬息万变的科学环境中(Stephan, 2012; Wang & Barabási, 2021),每年都有成千上万的新人进入科学界,关于男性和女性如何从科学界消失的传统假设可能需要仔细修正。本研究的目的是勾勒出一些尝试性的一般性答案和下一步研究的可能方向。

五、学术职业研究中大数据方法运用面临的挑战与局限

使用 Scopus 类型的原始结构化大数据时面临的主要挑战与该数据无法明确确定科研人员的性别、年龄和学科有关。

(一) 第一个挑战: 性别确定

科学家的性别不会直接出现在出版物中,但可以根据出版物中出现的几个变量来定义。在全球尺度的分析上,科学家的出版物是确定科学家性别的唯一数据来源,而在单个国家尺度的分析上,科学家的国家登记册可能是可用的。但是,从出版物数据中推断性别是有局限性的。简而言之,文献计量数据库并不完美,历史上也不是为研究科学职业或任何其他研究课题而设计的(因此有必要将这类数据重新用于研究目的;可参考 Salganik, 2018)。文献计量数据库的标准重点是出版物;同样,大规模科学研究的传统分析单位是出版物,而不是作者。划分科学家性别的三种主要方式之间存在着本质区别。

1. **行政数据:** 在国家科学家登记册和列出国家科学工作者的其他官方数据集中,通过行政方式收集性别信息,并标明科学家及其国家级 ID。

2. **自我报告:** 在调查研究(传统上是二元的,但在过去十年中越来越多地超越了二元方法)和基于访谈的研究中自我申报的性别。

3. **基于全球出版数据集的性别检测工具:** 性别由基于全球出版数据集元数据的性别检测工具确定,不同国家的准确度不同。

本研究采用了第三种性别标记方法。一方面,某些国家和某些国家数据集(意大利、挪威、波兰和美国就是最好的例子,这些国家和数据集对科学家进行了直接的、行政性的和二元的性别区分)所使用的完美的二元性别识别方法与跨国或全球研究中使用的性别区分方法之间形成了鲜明的对比。行政标记的性别准确概率为 100%,基本无误差;而本研究使用的大型性别检测工具的准确概率至少为 85%,存在一定误差。

探索过去三十年中女性对科学发展的贡献需要一个关键的判断:首先,需将科学家明确地归入两个性别组别中,即男性科学家或女性科学家。此外,还需要将科学家明确归入一个学科领域、所属国家和年龄组(以 Scopus 数据库中首次收录出版物的年份为基础)。利用单个研究人员(而不是单个出版物)开展大规模研究的主要挑战正是研究人员的性别。根据按国家、学科和时间划分的(编入索引的)出版物数量的增长情况来研究科学的发展,在方法和技术上似乎不那么复杂,因为这些指标构成了 WoS 或 Scopus 类型的大规模的、经过整理的商业数据集的核心。更复杂的是基于这些数据库中不断增加的研究人员(即出版物作者)来开展研究。在方法论上更为复杂的是那些为分析国家、学科、时间和性别层面科学家人数增长情况而进行的研究。本研究正是从后者的角度探讨了科学的发展,在每位科学家的数据组合中增加两个额外的维度:学术年龄和单一的主导科学学科。

基于文献计量数据库进行性别推断存在以下局限性:首先,并非所有数据库收录的出版物都包含作者的全名(这是性别检测工具识别性别的关键),尤其是天文学等学科的早期出版物和多作者出版物;其次,有全名的出版物比例因学科、研究时期和国家的不同而存在差异;再次,不同国家存在不分性别名字(Andrea, Claude & Dominique; Sabo, 2021),这增加了性别检测的难度。此外,目前使用的性

别检测工具的准确度也因所研究的国家和获取服务的成本而异。大多数工具是商业性质的,但也有一些免费或收取少量费用的工具可供选择。

此外,使用性别检测工具还涉及伦理问题:“性别”(sex)与“性别”(gender)的概念不可互换(生物学特征与社会文化作用),而性别检测工具通常采用二元性别分类。尽管可能通过学术专业调查和访谈中的自我报告,以及手动浏览科学家的机构网页内容来识别非二元性别,但大规模文献计量数据集无法识别非二元性别(Peters & Norton, 2018)。

在科学和科学家研究领域,性别检测工具的出现之前,性别鉴定主要依赖人工方式。然而,人工性别鉴定并非一种可扩展的方法,它通常基于对特定国家姓名的了解(Santamaría & Mihaljević, 2018)。此外,人工性别鉴定还可以借助社会保障管理部门的婴儿姓名数据库,或采用无监督的性别识别方法,例如将基于姓名的检测方法与基于图像的人脸识别技术相结合。

Karimi 等(2016)利用一个包含 1 400 多名科学家的简历、网站和面部照片的数据集,对几种可用的检测工具进行了比较。Wais(2016)介绍了一种基于 genderize.io 性别检测工具的 R 软件包,并比较了两项具有影响力的大规模女性科学家研究(Larivière et al., 2013; West et al., 2013)中使用的方法。Hu 等(2021)开发并比较了性别检测的监督学习方法,他们在一家互联网公司(Verizon Media)的注册用户数据集上训练了机器学习模型,该数据集包含 2,100 万个唯一的性别/姓名对。他们还在该数据集与包含 98,400 个唯一姓名/性别对的社会安全局婴儿姓名数据集上评估了自己的模型。最近,大型语言模型(LLMs)在性别检测方面的潜力也得到了测试,多项研究表明,ChatGPT 的性能优于其他工具(Alexopoulos et al., 2023)。

在估算科学界女性比例的变化时,主要面临两大误差类型:测量误差和抽样误差。首先,测量误差可能源于欧洲样本中确定每位科学家性别的方法不够准确。本研究采用了 NamSor 这一可靠的性别检测工具,它内置了性别确定准确度的测量方法。NamSor API 返回的结果能够为给定的姓名组合提供最有可能的性别,并附带相应的统计数据。其次,抽样误差源于女性在科学家人口中所占比例的估算,是基于可以推断出性别的科学家样本得出的。因此,欧洲科学家样本并非随机抽取,而是一个非概率样本。样本的纳入取决于全名的可用性以及 NamSor 根据全名识别性别的能力。这意味着所使用的样本可能无法代表整个欧洲科学家群体,从而引入了抽样误差。不过,对于大多数国家而言,姓名性别模糊的科学家比例较低,因此潜在在抽样误差的影响较小。在本文中,按照 Science-Metrix(2018)的观点,我们假设这种偏差可以忽略不计,并且我们的样本对姓名含糊的作者具有代表性。对来自不同国家、学科和年龄段的数千名科学家个体进行人工整理的全面验证(基于分层随机抽样)将非常有用,但这超出了我们研究团队的能力范围。

本研究沿用了爱思唯尔近期关于女性在科学中作用的报告(Elsevier 2018; Elsevier 2020)中使用的方法。国际研究中心(ICSR)实验室向我们提供了有关科学家特定性别的数据。我们使用的数据集包含研究人员在 Scopus 数据库中的标识符和他们的性别,概率分数为 85%。我们使用的数据集包含 Scopus 数据库中的作者标识符和使用 NamSor 性别推断工具确定的两个变量。NamSor 具有很高的准确率(即误报率很低)和召回率(即未知数很少)以及全球覆盖率。其验证程序依赖于目录中列出的名称和地理位置(Science-Metrix 2018; NamSor 2024)。我们数据集中用于性别检测的 NamSor 软件在许多研究中都得到了积极评价(例如 Santamaría & Mihaljević 2018)。

(二) 第二个挑战: 确定学术年龄

在估算科学界女性比例的变化时,年龄标记同样面临挑战,主要体现在以下三种方法的区别上:一是国家科学家登记册中的年龄数据。在国家科学家登记册中,行政分配的出生年份与个人数据一致,可直接确定任何时间点的生理年龄。二是调查和访谈中的自我申报年龄。以调查和访谈为基础的研究中自我申报的生理年龄,有一定的误差范围。三是学术年龄的确定。学术年龄是通过全球出版数据集确定的,该数据集只能通过被数据库收录的首次发表论文的时间来确定学术年龄,不同国家和学科

的准确度不同。

本研究采用了第三种方法进行年龄标记。正如学术职业的传统研究(Stephan & Levin, 1992)和最近的科学经济学研究(Stephan, 2012)所强调的,年龄对学术职业的许多方面都很重要,包括研究生产力模式、出版模式以及移民和国际合作模式。通过首次发表论文的日期来确定学术年龄虽然存在一定的局限性,但这种方法在跨国研究中具有较高的可操作性和一致性。

在成熟的科学体系中,严格确定学术年龄的想法非常奏效,因为这些体系中的科学家在全球科学期刊中已经出现了几十年;而对于全球科学事业的新成员来说,这种想法并不奏效,因为他们过去一二十年中才越来越多地出现在数据集中。因此,某些国家和一些国家科学家登记处使用的完美年龄识别方法,与本研究及许多基于文献计量学的研究中使用的(学术)年龄作为生物年龄的替代方法,形成了鲜明的对比。行政分配的出生年份(无差错)与根据首次索引出版物的元数据得出的(学术)年龄形成鲜明对比。学术年龄的确定在不同的系统中有不同的准确度,这取决于国家系统在全球索引出版物系统中的嵌入程度。

确定年龄的方法问题对于两类科学家分析至关重要,而大数据使得以下分析成为可能:一是队列分析,可以对不同时期开始发表论文的科学家或属于不同出版队列的科学家进行比较;二是纵向分析,可以对发表论文的科学家进行长期跟踪,从他们第一篇论文被数据集收录开始。

自21世纪初开始对个人进行微观层面的大规模研究以来,学术职业研究中越来越多地使用学术年龄来替代生理年龄(Nane et al. 2017; Robinson-Garcia et al. 2020)。大规模研究(相对于小规模研究)往往无法获得生理年龄数据以用于研究目的。半个多世纪以来,生理年龄一直被用于科学和科学家的研究,从雷曼(Lehman)的《年龄与成就》(1953)一书开始,生理年龄就受到了广泛的批评;还有一系列研究侧重于生产力对科学界社会分层的影响,也使用了年龄数据,如Pelz和Andrews(1976);Kyvik(1990);以及Stephan和Levin(1992)。

在科学界,确定科学家的学术年龄和生理年龄之间的关系是一个复杂的问题。传统上,生理年龄的替代指标主要有两种:第一,与首次发表论文日期相关的学术年龄(如Nane et al., 2017; Radicchi & Castellano, 2013; Robinson-Garcia et al., 2020);第二,与获得博士学位日期相关的学术年龄(van den Besselaar & Sandström, 2016; Savage & Olejniczak, 2021)。在单个国家研究中,这两种方法均适用,但在跨国研究中,基于博士学位日期的学术年龄更具可操作性。随着研究范围从国内扩展到多国或全球,生理年龄通常只能通过学术年龄的替代值来估算。

系统分析表明,学术年龄和实际生理年龄之间的相关性在不同学科之间存在显著差异(Kwiek & Roszka, 2022)。在本研究中,对于STEMM学科,相关系数通常在80%-90%之间(例如化学为0.89,免疫学和微生物学为0.90)。对于科学相对“落后”的国家(本研究未涉及),相关性可能更为显著,而对于“成熟”和“发展中”国家,相关性可能较小。

使用基于Scopus的学术年龄作为生理年龄的替代指标,虽然存在数据集偏差的问题,但Scopus在跨国研究中具有明显优势。与Web of Science相比,Scopus在覆盖范围和数据完整性方面更为有效。其他潜在的数据源包括谷歌学术数据库(Google Scholar)和OpenAlex。谷歌学术数据库偏向于年轻科学家(Radicchi & Castellano, 2013),而OpenAlex在处理具有明确定义的出版和引用组合的最底层个人数据时存在困难(Priem et al., 2022)。

所有数据来源仅涵盖有发表记录的科学家,而无发表记录的科学家未被纳入分析。这意味着,女性(和男性)参与科学发展的情况完全是通过科研发表的视角来观察的,其他角色(如教学、行政管理和指导等)均被排除在外。这种局限性需要在解读研究结果时予以考虑。

(三) 第三个挑战: 确定学科

在确定研究样本中每位科学家的学科时,我们面临多种方法选择。在国家数据集中,科学家的学

科划分通常依据国家研究评估工作常用的分类标准, 科学家一般被归入低阶学科和高阶领域。

当从国家研究转向多国研究时, 主要挑战在于如何在 32 个国家的学科清单中找到共同标准。鉴于协调 32 种不同分类的难度, 本研究选择采纳 Scopus 数据库的广泛领域和学科分类。样本中所有科学家的出版物最初均被归入期刊, 并标注了 ASJC 学科代码。此外, 所有出版物都包含引用的参考文献列表。通过整合每位科学家的出版物和引用参考文献, 我们确定了每位科学家的单一主导学科。

具体而言, 本研究使用了 Scopus 数据库中的一组出版物, 这些出版物发表于 2023 年及之前, 并限于期刊论文和会议论文。每篇出版物中的每条被引用参考文献都附有其所在期刊的学科信息, 即四位 ASJC 代码。为了转换为两位数分类法, 本研究获取了四位数代码的前两位。随后, 统计每位作者引用各学科的次数, 并选择被引用次数最多的学科作为主导学科。我们剔除了有多个主要学科或无学科的作者。

对于发表时间较长、引用较多的科学家, 这种方法通常更有效; 而对于发表时间短、产出有限的科学家, 学科归属可能更具主观性。基于至少 10 篇文章的学科归属可能更准确, 但这会排除许多早期职业研究人员或已离开科学界的科学家 (Kwiek & Szymula, 2024)。本研究采用了“一个科学家, 一个学科”的简化方法, 尽管科学家可能会在职业生涯中更换学科。

(四) 研究局限

本研究在探讨科学职业发展时存在一定的局限性, 主要体现在以下几个方面:

研究对象的局限性: 本研究仅关注发表论文的科学家, 未涵盖未发表论文或发表论文水平较低的科学家。研究对象为各领域的发表论文科学家, 而非所有高等教育人员。

“离开科学界”的概念化: 本研究中“离开科学界”被定义为停止发表工作, 但这一概念并未涉及其他学术角色 (如教学、行政管理) 或非学术角色 (如在企业或政府工作)。科学界与非科学界工作之间的界限较为模糊 (Preston, 2004, p. 14), 使得“离开科学界”成为一个难以精确界定的概念。

学术生命周期的界定: 本研究通过科学家在 Scopus 数据库中发表的第一篇和最后一篇学术论文来探讨其学术生命周期。这种方法将发表论文的序列视为从事科学工作的主要指标, 假设未在学术期刊上发表论文即意味着不再从事科学工作 (Sugimoto & Larivière, 2023)。因此, 本研究对科学职业的描述简化为出版生涯, 对学术成果的表述简化为全球收录的出版物。

数据集的局限性: 本研究中每位科学家的个人终身出版物组合仅限于被 Scopus 数据库收录的出版物, 未被收录的出版物 (尤其是非英语出版物) 未被计算在内。此外, 科学家在学术界的其他活动 (如指导学生、评审论文、审核资助提案和编辑期刊等) 在研究中被忽略 (Liu et al., 2023)。换言之, 在本研究中, 积极参与科学研究主要通过出版行为来体现。根据科学社会学的默顿传统, 不出版被定义为“离开科学界”。尽管 STEMM 领域的非出版者可能继续在学术部门工作或担任其他学术职务, 但利用文献计量数据集无法在全球范围内核实他们的具体就业情况。

(马瑞克·科维克工作邮箱: kwiekm@amu.edu.pl. 译者: 孙娜, 华东师范大学高等教育研究所博士生)

参考文献

- Alexopoulos, M., Lyons, K., Mahetaji, K., Barnes, M. E., & Gutwillinger, R. (2023). Gender inference: can chatGPT outperform common commercial tools?. *arXiv preprint arXiv:2312.00805*.
- Allison, P. D. (2014). *Event history and survival analysis*. Sage.
- Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1), 377—386.
- Blickenstaff, J. C. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, 17(4), 369—386.
- Branch, E. H. ed. (2016). *Pathways, Potholes, and the Persistence of Women in Science. Reconsidering the Pipeline*. Lexington Books.
- Branch, E. H. , & Alegria, S. (2016). Gendered Responses to Failure in Undergraduate Computing. Evidence, Contradictions, and New Direc-

- tions. In: E. H. Branch (ed.), *Pathways, Potholes, and the Persistence of Women in Science. Reconsidering the Pipeline* (p. 17-31). Lexington Books.
- Cornelius, R., Constantinople, A., & Gray, J. (1988). The chilly climate: Fact or artifact? *The Journal of Higher Education*, 59(5), 527—55.
- Deutsch, F. M., & Yao, B. (2014). Gender differences in faculty attrition in the USA. *Community, Work & Family*, 17(4), 392—408.
- Ehrenberg, R. G., Kasper, H., & Rees, D. I. (1991). Faculty turnover in American colleges and universities. *Economics of Education Review*, 10(2), 99—110.
- Elsevier (2018). *Gender in the global research landscape*. Elsevier.
- Elsevier. (2020). *The researcher journey through a gender lens*. Elsevier.
- Fox, M. F. (2010). Women and Men Faculty in Academic Science and Engineering: Social-Organizational Indicators and Implications. *American Behavioral Scientist*, 53, 997—1012.
- Fox, M. F., & Mohapatra, S. (2007). Social-organizational characteristics of work and publication productivity among academic scientists in doctoral-granting departments. *The Journal of Higher Education*, 78(5), 542—571.
- Fox, M. F. (2020). Gender, science, and academic rank: Key issues and approaches. *Quantitative Science Studies*, 1(3), 1001—1006.
- Fox, M. F., & Xiao, W. (2013). Perceived Chances for Promotion Among Women Associate Professors in Computing: Individual, Departmental, and Entrepreneurial Factors. *Journal of Technology Transfer* 38: 135—152.
- Fox, M. F., & Kline, K. (2016). Women Faculty in Computing. A Key Case of Women in Science. In: E. H. Branch (ed.), *Pathways, Potholes, and the Persistence of Women in Science. Reconsidering the Pipeline* (p. 54-69). Lexington Books.
- Geuna, A., & Shibayama, S. (2015). Moving out of academic research: Why do scientists stop doing research? In A. Geuna (Ed.), *Global mobility of research scientists* (pp. 271—297). Elsevier.
- Glenn, Norval D. 2005. *Cohort analysis*. Thousand Oaks, CA: Sage
- Goulden, M., Mason, M. A., & Frasch, K. (2011). Keeping women in the science pipeline. *The ANNALS of the American Academy of Political and Social Science*, 638, 141—162.
- Hammarfelt, B. (2017). Recognition and reward in the academy: Valuing publication oeuvres in biomedicine, economics and history. *Aslib Journal of Information Management*, 69(5), 607—623.
- Hu, Y., Hu, C., Tran, T., Kasturi, T., Joseph, E., & Gillingham, M. (2021). What's in a name?—gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 35(4), 1537—1563.
- Kaminski, D., & Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science*, 335, 864—866.
- Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology*, 82(5), 965—990.
- Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web* (pp. 53-54).
- Kashyap, R., Rinderknecht, R. G., Akbaritabar, A., Albrez-Gutierrez, D., Gil-Clavel, S., Grow, A., Zhao, X. (2023). Digital and computational demography. In: *Research handbook on digital sociology*, 48—86. Cheltenham: Edward Elgar Publishing.
- King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, 3.
- Kwiek, M. (2016). The European research elite: A cross-national study of highly productive academics across 11 European systems. *Higher Education*, 71(3), 379—397.
- Kwiek, M. (2018). High research productivity in vertically undifferentiated higher education systems: Who are the top performers? *Scientometrics*. Vol. 115(1). 415—462.
- Kwiek, M., & Roszka, W. (2021a). Gender disparities in international research collaboration: A large-scale bibliometric study of 25, 000 university professors. *Journal of Economic Surveys*, 35(5), 1344—1388.
- Kwiek, M., & Roszka, W. (2021b). Gender-based homophily in research: A large-scale study of man-woman collaboration. *Journal of Informetrics*, 15(3), 1—38.
- Kwiek, M., & Roszka, W. (2022). Are female scientists less inclined to publish alone? The gender solo research gap. *Scientometrics*, 127, 1697—1735.
- Kwiek, M., & Szymula, L. (2024). Quantifying attrition in science: A cohort-based, longitudinal study of scientists in 38 OECD countries. *Higher Education*, 1—29. Online first: <https://doi.org/10.1007/s10734-024-01284-0>.
- Kyvik, S. (1990). Age and scientific productivity. Differences between fields of learning. *Higher Education*, 19(1), 37—55.
- Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504(7479),

- 211—213.
- Liu, L., Jones, B. F., Uzzi, B., et al. (2023). Data, measurement, and empirical methods in the science of science. *Nature Human Behaviour*, 7, 1046—1058.
- Loannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLOS One*, 9(7), e101698.
- Maranto, C. L., & Griffin, A. E. (2011). The antecedents of a “chilly climate” for women faculty in higher education. *Human Relations*, 64(2), 139—159.
- Mills, M. (2011). *Introducing survival and event history analysis*. Sage.
- Milojevic, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences*, 115, 12616—12623.
- NamSor (2024). NamSor API documentation available from <https://namsor.app/api-documentation/>.
- Nane, G. F., Larivière, V., & Costas, R. (2017). Predicting the age of researchers using bibliometric data. *Journal of informetrics*, 11(3), 713—729.
- Nielsen, M. W., & Andersen, J. P. (2021). Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, 118(7), e2012208118.
- Nygaard, L. P., Piro, F., & Aksnes, D. (2022). Gendering excellence through research productivity indicators. *Gender and Education*, 34(6), 690—704.
- Pelz, D. C., & Andrews, F. M. (1976). *Scientists in organizations: Productive climates for research and development*. New York: Wiley.
- Peters, S. A., & Norton, R. (2018). Sex and gender reporting in global health: new editorial policies. *BMJ global health*, 3(4).
- Preston, A. E. (2004). *Leaving science. Occupational exit from scientific careers*. Russell Sage Foundation.
- Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
- Radicchi, F., & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, 97, 627—637.
- Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V., & Nane, G. F. (2020). Task specialization across research careers. *Elife*, 9, e60586.
- Rosser, V. J. (2004). Faculty members’ intentions to leave: A national study on their work-life and satisfaction. *Research in Higher Education*, 45(3), 285—309.
- Santos, J. M., Horta, H., & Amâncio, L. (2020). Research agendas of female and male academics: A new perspective on gender disparities in academia. *Gender and Education*, 1—19.
- Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4, e156.
- Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press.
- Savage, W. E., & Olejniczak, A. J. (2021). Do senior faculty members produce fewer research publications than their younger colleagues? Evidence from Ph. D. granting institutions in the United States. *Scientometrics*, 126(6), 4659—4686.
- Science-Metrix (2018). *Analytical Support for Bibliometrics Indicators. Development of bibliometric indicators to measure women’s contribution to scientific publications. Final Report*. Science-Metrix.
- Shaw, A. K., & Stanton, D. E. (2012). Leaks in the pipeline: Separating demographic inertia from ongoing gender differences in academia. *Proceedings of the Royal Society B. Biological Sciences*, 279(1743), 3736—3741.
- Shibayama, S., & Baba, Y. (2015). Impact-oriented science policies and scientific publication practices: The case of life sciences in Japan. *Research Policy*, 44(4), 936—950.
- Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis. Modeling Change and Event Occurrence*. Oxford: Oxford University Press.
- Smart, J. C. (1990). A causal model of faculty turnover intentions. *Research in Higher Education*, 31(5), 405—424.
- Spoon K., LaBerge N., Wapman K. H., Zhang S., Morgan A. C., Galesic M., Fosdick B. K., Larremore D. B., Clauset A. (2023). Gender and retention patterns among U. S. faculty. *Science Advances*, 9, eadi2205.
- Stephan, P. E., & Levin, S. G. (1992). *Striking the mother lode in science: The importance of age, place, and time*. New York: Oxford University Press.
- Stephan, P. E. (2012). *How economics shapes science*. Harvard University Press.
- Sugimoto, C., & Larivière, V. (2023). *Equity for women in science: Dismantling systemic barriers to advancement*. Harvard University Press.
- Tang, L., & Horta, H. (2023). Supporting academic women’s careers: Male and female academics’ perspectives at a Chinese research university. *Minerva*. <https://doi.org/10.1007/s11024-023-09506-y>.

- Van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: a longitudinal case study. *Scientometrics*, 106, 143—162.
- Wang, D., & Barabási, A. -L. (2021). *The science of science*. Cambridge University Press.
- Wais, K. (2016). Gender prediction methods based on first names with genderizeR. *The R Journal* 8(1): 17—37.
- West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PloS one*, 8(7), e66212.
- White-Lewis, D. K., O'Meara, K., Mathews, K., et al. (2023). Leaving the institution or leaving the academy? Analyzing the factors that faculty weigh in actual departure decisions. *Research in Higher Education*, 64, 473—494.
- Wohrer, V. (2014). To stay or to go? Narratives of early-stage sociologists about persisting in academia. *Higher Education Policy*, 27, 469—487.
- Xie, Y., & Shauman, K. A. (2003). *Women in Science: Career Processes and Outcomes*. Harvard University Press.
- Xu, Y. J. (2008). Gender disparity in STEM disciplines: A study of faculty attrition and turnover intentions. *Research in Higher Education*, 49, 607—624.
- Zhou, Y., & Volkwein, J. F. (2004). Examining the influence on faculty departure intentions: A comparison of tenured versus nontenured faculty at research universities using NSOPF-99. *Research in Higher Education*, 45(2), 139—176.

(责任编辑 王 森)

Patterns of Scientific Attrition: A Quantitative Analysis of Research Landscapes in OECD Countries with Methodological Considerations

Marek Kwiek¹ Lukasz Szymula²

(1. Center for Public Policy Studies, Adam Mickiewicz University of Poznan, Poznan 61712, Poland;

2. Faculty of Mathematics and Computer Science, Adam Mickiewicz University of Poznan, Poznan 61712, Poland)

Abstract: In this paper, we explore how members of the scientific community leave academic science and how attrition (defined as ceasing to publish) differs across genders, academic disciplines, and over time. Our approach is cohort-based and longitudinal. We track individual male and female scientists over time and quantify the phenomenon traditionally referred to as “leaving science.” Using publication metadata from Scopus—a global bibliometric database of publications and citations—we follow the details of the publishing careers of scientists from 38 OECD countries who started publishing in 2000 (N = 142,776) and 2010 (N = 232,843). Within a decade, about 50% of scientists stay in science and after 19 years – only about 30%. Our study is restricted to 16 STEMM disciplines (science, technology, engineering, mathematics, and medicine), and we track the individual scholarly output of the two cohorts until 2022. We use survival analysis to compare attrition of men and women scientists. With more women in science and more women within cohorts, attrition is becoming ever less gendered. In addition to the combined aggregated changes at the level of all STEMM disciplines, widely nuanced changes were found to occur at the discipline level and over time. Attrition in science means different things for men versus women depending on the discipline; moreover, it means different things for scientists from different cohorts entering the scientific workforce. Finally, global bibliometric datasets were tested in the current study, opening new opportunities to explore gender and disciplinary differences in attrition. It also describes the methodological challenges and research limitations of using raw structured big data in academic career research, such as identifying gender, age, and discipline.

Keywords: academic careers; leaving science; big data; survival analysis; science of science