·方法学·

UK Biobank 数据的应用介绍

黄夏璇[1,2#]，黄韬[2#]，杨瑞[2]，袁师其[1,2]，何宁霞[2]，徐安定[1]，吕军[2,3]

1. 暨南大学附属第一医院神经内科（广州 510630）
2. 暨南大学附属第一医院临床研究部（广州 510630）
3. 广东省中医药信息化重点实验室（广州 510632）

【摘要】 UK Biobank（英国生物银行）是一个大型生物医学数据库和研究资源，包含来自 50 万英国参与者的生物遗传和健康信息，涵盖了丰富的基本结构化数据、高通量的基因组学遗传数据和多模态影像数据，但是由于其数据体量大，使用方式复杂，因此国内研究者使用的并不广泛。本文首先介绍 UK Biobank 的健康相关结构数据、基因数据和影像数据等内容，随后分别对不同数据的下载使用方法进行详细描述，最后探索近几年使用 UK Biobank 数据库进行的最新研究，并围绕将人工智能技术应用于 UK Biobank 数据的典型研究和发展方向进行探讨，以期在人体解剖生理和遗传变异、表型特征等方面有更多的科学研究成果出现。

【关键词】 UK Biobank；数据库；影像；基因组学；人工智能

Introduction of UK Biobank data application

HUANG Xiaxuan[1,2], HUANG Tao[2], YANG Rui[2], YUAN Shiqi[1,2], HE Ningxia[2], XU Anding[1], LV Jun[2,3]

1. Department of Neurology, The First Affiliated Hospital of Jinan University, Guangzhou 510630, P. R. China
2. Department of Clinical Research, The First Affiliated Hospital of Jinan University, Guangzhou 510630, P. R. China
3. Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Guangzhou 510632, P. R. China
Corresponding author: XU Anding, Email: tlil@jnu.edu.cn; LV Jun, Email: lyujun2020@jnu.edu.cn

【Abstract】 UK Biobank is an extensive biomedical database and research resource. It contains in-depth genetic and health information from 500 000 UK subjects, comprising a wealth of basic structured data, high-throughput genomic and genetic data, and multimodal imaging data. However, difficulties in accessing the large amount of data mean that the database has not been widely used in China. We first introduced the health-related structural data, genetic data, and imaging data in the UK Biobank. We then described methods for using different types of data downloaded from UK Biobank, and explored recent research based on these data. We also discussed classic research focusing on applying artificial intelligence technology to UK Biobank data. Finally, we predicted future research trends in the utilization of UK Biobank data in areas such as anatomy, physiology, genetic variation, and phenotypic characteristics.

【Key words】 UK Biobank; Database; Imaging; Genomics; Artificial intelligence

1 引言

英国生物银行（UK Biobank，UKB）[1-2] 是一个大型的生物医学数据库和研究资源，作为一项前瞻性的流行病学科学研究计划，共收集了英国各地年龄在 40～69 岁之间的 50 万例志愿者数据信息，包括了志愿者的基因数据、多模态影像数据及健康相关数据。这项研究计划的时间跨度从 2006 年延续至今，并且官方表示在未来 30 年内，将长期追踪该人群的健康和医疗状况信息。直至 2020 年初，UKB 已处理并发布了超过 45 000 人的影像数据[3]，为后续更长远地从人体生理解剖和遗传基因表型方面探讨疾病的诊断和治疗提供了大数据基础。UKB 数据库不仅定期增加额外数据和更新，而且对全球符合有关伦理和科学标准的研究人员开放访问，促成多项改善人类健康新项目的产生，并得到了一些新的发现。

UKB 的研究领域不仅限于对人群的基因表型数据进行全基因组关联分析，随着多模态影像数据在神经系统方面的资源被深度开发[4]，研究人员开

DOI：10.7507/1672-2531.202204162
基金项目：广东省中医药信息化重点实验室项目（编号：2021B1212040007）
通信作者：徐安定，Email：tlil@jnu.edu.cn；吕军，Email：lyujun2020@jnu.edu.cn
#共同第一作者

始对大脑结构和功能，从行为学和临床结果等方面对疾病的预后预测及风险因素进行探索。另外，深度学习作为机器学习中最新的研究方向[5-6]，通过复杂的机器学习算法，主要目标是让机器能够像人一样具有分析学习能力，通过学习样本数据的内在规律和具体特征，进而能够识别目标数据的文字、图像和声音等信息。本研究将对 UKB 数据库进行基本介绍，分别阐述不同类型数据的使用方法，并围绕将人工智能技术应用于 UKB 数据的典型研究和未来发展方向进行探讨，期待更多的研究人员可从人体解剖生理角度和遗传变异、表型特征等方面，为与公共卫生密切相关的疾病预防和治疗开辟新的研究领域。

2　UKB 数据库整体介绍

UKB（数据库官网：https://www.ukbiobank.ac.uk/）是全球最大的生物医学样本数据库，也是世界上最详细、最长期的前瞻性健康研究。在 2006—2010 年间，UKB 作为一项纵向研究，从英国各地招募了 50 万例年龄在 40 ~ 69 岁的志愿者，计划收集大约 1 500 万份血液、尿液和唾液的生物样本，并对所有参与者进行基因分型和血液生化分析，调查志愿者的生活方式（包括营养、生活方式和药物使用情况等）及亲属遗传关系，长期追踪他们的健康和医疗状况信息，并要求每位参与者都参加在英格兰或者苏格兰、威尔士的中心医院进行的基线评估。另外，UKB 的影像扩展项目[7] 于 2016 年获得资助，计划到 2023 年初步完成，该项目拟扫描 100 000 个现有 UKB 队列对象，包括对大脑、心脏和身体的磁共振成像检查（magnetic resonance imaging，MRI）、骨和关节低剂量 X 射线扫描及颈动脉超声检查，扫描成像队列中所有受试者的影像数据采集在 3 个专业影像检查中心完成。从 2017 年 6 月公开至今，UKB 数据库收集并供使用的开放数据主要包括：所有参与者的健康相关数据（死亡数据、癌症数据、初级保健记录、住院记录数据）、生化样本分析、物理活动检测、问卷调查、基线评估数据、多模态成像、全基因组基因分型的纵向随访数据。图 1 展示了不同数据的开放时间和先后次序。目前，UKB 仍在持续不断地更新，表 1 展示了 UKB 未来 2 年内的数据发布细节和时间。另外，需要访问该研究资源的研究人员必须在访问管理系统（access management system，AMS）中填写注册表在英国生物银行注册，并且通过 AMS 系统申请访问数据库，填写个人研究摘要及所需数据内容，待英国生物银行批准审核完成后方可使用 UKB 的部分数据内容。总的来说，UKB数据库不同于其他数据库的特点就是，围绕健康人群为主、具有丰富的样本量和数据并且更新速度快。

3　UKB 数据库分布

3.1　基线评估

在 2006—2010 年从英国招募了 40 ~ 69 岁之间的健康人群，在苏格兰、英格兰和威尔士的 22 个评估中心进行了基线评估，主要包括书面同意、饮食回忆记录、肺活量和骨密度测量情况、血液、尿液和唾液样本采集等内容[8]，其中，血液数据覆盖患者从贫血到血源性癌症等血液疾病，和与癌症及其他慢性非传染性疾病相关的 20 种病原体血清学抗体反应，以及进行基于核磁共振成像（nuclear magnetic resonance，NMR）的代谢组学测定的 200 多种代谢产物，这些数据在未来几年将持续提供及更新（具体内容详见：https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/baseline-assessment）。

3.2　在线问卷

英国生物银行定期向 330 000 例参与者进行问卷调查，发送电子邮件地址，每份问卷的回复率达 35% ~ 50%，收集关于 24 小时饮食回忆、认知功能、职业经历、终身和当前的心理健康、消化系统健康、慢性疼痛及食物偏好，未来计划将对睡眠情况、神经发育情况及生活质量。

3.3　基因数据库

1999 年提议建设的 UKB 数据库研究计划，旨在建成世界上最大的有关致病或预防疾病的基因信息库。从 2017 年 7 月开始更新基因数据，对所有英国生物银行参与者进行全基因组基因分型、全外显子组测序和全基因组测序，将大大改变研究人员研究各种健康结果的遗传学决定因素的方式[9-10]。其中，英国生物样本库的遗传数据包含了 488 377 例参与者的基因型，同时提供了人类白细胞抗原区（human leukocyfe antigen，HLA）的各种基因排列数据运算。据报道，UKB 数据库利用自主设计的基因分型芯片对 50 万被试者进行全基因组单核苷酸多态性（single nucleotide polymorphism，SNP）数据搜集，并且自主开发了一套针对 UKB 数据的管理系统，从规模、多样性及特异性等特点对其收集的基因数据进行质量控制，全部基因数据包括了 50 万人的 9 600 万个位点的基因变异信息[11]。现如今大部分的研究人员使用一种基于阵列的方法来

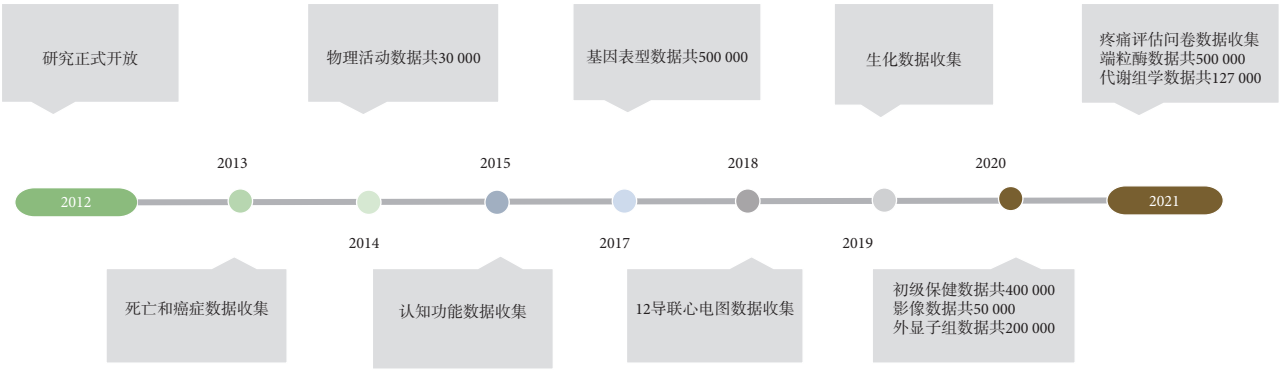

图 1　UKB 数据开放时间

表 1　UKB 未来数据的发布情况

数据类型	细节	发布日期
SARS-CoV-2血清学数据	2020年6月至2020年12月和2021年11月至2022年2月期间20 000人（10 000例英国生物银行参与者及其10 000例成年子女/孙辈）的SARS-CoV-2抗体状态数据	2022年第4季度
苏格兰中心癌症数据	苏格兰癌症登记记录的更新，涵盖2015年10月以后的时期	2022年第4季度
营养数据	已计算已完成饮食网络问卷的参与者的最新营养信息和总膳食摄入量	2022年第4季度
医院住院数据（威尔士）	威尔士住院数据的更新，涵盖自2018年2月以后的时期	2022年第4季度
死亡登记数据	在数据门户和展示字段中更新英格兰，苏格兰和威尔士的死亡登记数据	2022年第4季度
新冠肺炎测试结果数据	数据门户上英格兰，苏格兰和威尔士的新冠肺炎测试结果数据	2022年第4季度
住宅位置数据	开始改进和更新目录第150条目中现有的住宅位置数据	2022年第4季度
苏格兰精神病学数据	更新苏格兰的住院数据，以包括精神病入院	2022年第4季度
认知功能问卷数据	来自超过170 000例参与者的基于网络的认知功能问卷的数据	2022年第4季度
增强的癌症数据	有关癌症分期和分级的数据，以及癌症治疗，包括放疗、化疗周期和手术	2023年
南丁格尔第二阶段数据	Nightingale Health在220类中产生的现有NMR代谢组学数据将进行更新，以涵盖其他参与者，数据涵盖项目结束时预期的整个队列	2023年
WGS主要全基因组发布	覆盖整个队列的个体水平和联合命名的全基因组测序数据预计将于2023年最终发布	2023年

SARS-CoV-2：严重急性呼吸综合征冠状病毒2；WGS：全基因组测序。

确定基因数据的应用，收集遗传数据的特制基因分型矩阵，进行基因的分型和估算遗传相关性，可对UKB 基因数据库记录的全部性状与单个遗传变异之间的关联进行分析。目前主要涉及 2 个方面的研究，一方面，将基因数据进行优化，提取基因中的变异数据推算基因型和疾病的关系，了解疾病本身的生物学基础、遗传因素和生活方式因素之间的相互作用及疾病的潜在遗传学特征[12]。另一方面，将 UKB 的基因数据和影像数据进行结合分析，通过对不同的影像结构功能特征指标进行全基因组关联研究，观察遗传变异和成像特征之间的关联集群，得到基因与疾病之间的相关性，从而在疾病的发生机制中发现更多的遗传影响因素[13]。

3.4　影像数据库

自 2011 年起，UKB 成立了一个专家成像工作组，在与全球 100 多名成像专家协商后，开发了一种大规模的影像成像采集协议，旨在最大限度地提高收集成像数据的科学价值，同时也可在较短采集时间得到大规模实现。2014 年，UKB 启动了一项新的医疗成像数据收集计划，使用 MRI 和 X 射线技术对超过 10 万例志愿者进行分析，该项目包括大脑、心脏和身体的 MRI，骨骼和关节的全身双能 X 线吸收测定法（dule energy X-ray absorptiometry，DEXA）扫描，颈动脉的超声扫描，以及视网膜的光学相干断层扫描成像。图 2 展示了 UKB 此次项目所收集的影像数据的内容[3-4]。截至 2020 年初，超过 45 000 例参与者接受了评估，已经使 UKB 成像增强计划成为迄今为止世界上最大的多模态成像研究，其中已经有 10 000 例在第一次检查后 2 年返回进行重复成像，成像采集主要以多模态为主。所谓的多模态包括：3 种模态的结构 MRI 数据，静息态、任务态 fMRI 数据及 diffusion MRI 数据。UKB 提供了对影像进行全自动处理的流程使得影像在不同模态与样本间是可比较的，基于处理后的多模态影像数据生成上千个影像指标（imaging-derived phenotypes，IDPs）[14-15] 来描述人体解剖器官的结构与功能，多模态数据是直接利用统一的硬件和软件直接获取反映所有影像特征的多模态 IDPs

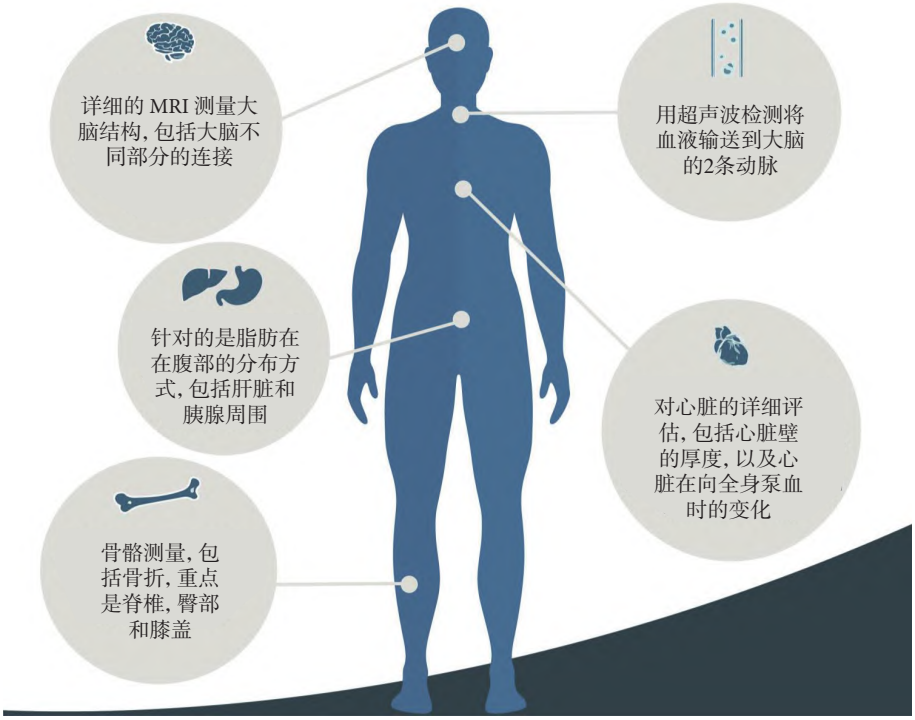

图 2　UKB 影像数据内容

指标。据不完全统计，截至 2020 年已经有超过 1 750 个关于 UKB 的研究持续进行中，也被不断用于解决一系列新的研究难题。与其他数据库不同的是，UKB 是首个采用复杂的全自动图像处理流程，提取了 4 350 个反映大脑不同结构功能特征的 IDPs，并且在核磁共振成像扫描中采用了最新的 MRI 采集技术，缩短了采集时间，大大提高了图像的空间分辨率[16-18]。UKB 影像数据库的一个独特特点就是大样本量人群及其多模态数据都是利用统一的硬件和软件获取，便于研究人员可在一定程度上不受分析和处理图像的繁杂流程的限制，大大提升了统计效率。

3.5　健康相关数据

健康相关数据包括了 UKB 数据库中各种与健康有关的电子记录数据[19]，主要有死亡、癌症、住院和初级保健记录数据，所有数据都在持续更新和随访，具体展示如下：① 死亡数据：UKB 通过与国家死亡登记处的联系定期收到死亡通知，每个参与者的主要死因由 ICD-10 代码确定，自项目开放以来，已经有 37 733 例参与者死亡，平均死亡年龄约为 69.6 岁，男性占平均死亡人数的 59%，其中主要死因是缺血性心脏病。② 癌症数据：UKB 囊括了最常见的癌症诊断，其中 25 503 例参与者被诊断为癌症，且诊断平均年龄为 52.2 岁，最常见的癌症为乳腺癌、前列腺癌和结直肠癌。③ 住院记录数据：医院住院患者主要来自英格兰（89%）、苏格兰（7%）和威尔士（4%）的医院住院数据，分别由数据库存取及检索系统（database access and retrieval system，DARS）、安全匿名信息链接数据库（secure anonymize information linkage，SAIL）、电子数据研究与创新服务数据库（electronic data research and innovation service，eDRIS）等不同的数据源收集而来，其中，医院住院患者数据中的所有临床数据都根据世界卫生组织的 ICD（国际疾病和相关健康问题分类）进行编码，所有的操作和程序都根据 OPCS（人口、人口普查和调查办公室：干预措施和程序的分类）进行编码。所有英国生物样本库关联的英格兰和大多数威尔士医院数据都用 ICD-10 和 OPCS-4 编码。然而，由于苏格兰数据的收集开始于更早的时间（1981 年），早期的苏格兰住院数据（1997 年之前收集的数据）采用 ICD-9 和 OPCS-3 编码，只有少量的威尔士住院记录用 ICD-9 编码，所有的电子病历数据一起收集了关于诊断和症状的类型和日期、程序和操作、处方、检测结果和全科医生转诊的信息。④ 初级保健记录数据：从 2019 年到现在，共 230 000 例参与者，主要包括：诊断、实验室检查、处方药、处方日期，药物代码、药物名称及数量等，并且定期提供最新的初级保健数据，甚至包括新型冠状病毒肺炎（COVID-19）相关的数据可供研究，但须遵守患者信息控制法规。

3.6　生化标志物

UKB 将所有 500 000 例参与者及 2012—2013

年参加重复评估访问的 20 000 例参与者中收集的样本中测量广泛的生化检查标志物。根据实验室检测度研究不同疾病的科学相关性，以血细胞计数（从所有参与者收集的新鲜血液样本的血液学检测）、传染病标志物（测量 10 000 例参与者针对 20 种病原体的血清学抗体反应）、代谢组学（从所有参与者收集的血液样本进行 NMR 代谢组学测定 200 多种代谢产物）、端粒长度（从所有参与者收集的血液样本提取 DNA 测量的染色体标志物）为具体分类，总共纳入 34 种生化标志物，包括：临床上已确定的疾病危险因素、诊断相关未明确的因素或未得到良好评估的表型标志物。

3.7 活动检测数据

在 2013 年 6 月至 2016 年 1 月之间，UKB 通过腕带式活动检测器收集 100 000 例参与者的 7 天内的体力活动数据，主要针对个体身体运动活动的测量情况，并被要求每季度需重复 4 次检查，包括听力和动脉僵硬测试、心肺健康测试、各种视力测量及佩戴加速计收集 7 天的身体活动数据等。

4 UKB 数据提取

UKB 的数据提取较为复杂，基本步骤如表 2 所示。主要分为 5 个步骤，一是获取数据校验码及密钥文件，二是下载个人项目数据包，三是检查数据完整性，四是解密解压，最后才能使用不同的工具下载相应的数据。

需要注意的是，只有经过申请，并通过官方授权的项目相关数据才可提取出来，没有通过授权的数据是无法下载的。授权密钥每年更新一次，提取数据时需把.key 文件放置在与数据包、提取工具相同的目录内。密钥除了规定哪些数据可下载外，还约束了数据的键值，不同项目的密钥与数据键值不同。接下来本章主要介绍几种不同数据类型的提取方式。

4.1 主数据提取

UKB 中的主数据集是由结构化数据组成。第二章提到的健康相关数据、生化标志物、活动检测、基线评估等都属于这一部分，也包含有一些影像相关或者基因相关的指标。通俗的说，只要是能够使用表格统计展示的数据，基本都属于这一部分。使用 ukbconv 工具提取，需要搭配.enc_ukb 后缀的数据包文件和.key 后缀的密钥文件使用，需放置在同一文件夹内。以 windows 系统下载数据为例：

下载单个指标命令：./ukbconv ukb45434.enc_ukb csv -s100021。其中 100021 为具体的指标，表示维生素 D。提取的数据如表 3 所示，其中 eid 为患者 ID（也即是数据键值），5 个列的数据均是维生素 D，只是其上线周期不同，具体参考官方说明（https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=100021）。将编号替换成相应的指标代码可下载自己需要的数据。

下载批量指标命令：./ukbconv ukb45434.enc_ukb csv -iid.txt 。其中 id.txt 是一个包含有多个指标代码的文本文件，文件内指标代码采用换行输入。将得到与单个指标相同的数据文件，不同的指标在文件中往列的方向扩展。

CSV 是研究人员常用的格式之一，同样可导出其他格式，如 bulk、docs、lims、r、sas、stata、txt 等，只需将命令中的 csv 替换成相应的字符串即可。

4.2 大批量数据提取

大批量的数据主要包括身体各部位的 MRI 影像、超声数据和 ECG 数据等。与提取主数据类似，需要将下载工具、.enc_ukb 后缀的数据包文件和.key 后缀的密钥文件放置在同一文件夹内（密钥文件建议改名为.ukbkey 全名），使用 ukbconv 和 ukbfetch 工具提取。不同的是，提取大批量数据不像提取主数据那么直接，包括 3 个步骤：① 检查电脑网络，能打开 UKB 官方的数据存储库网址 biota. ndph.ox. ac. uk 或 biota. ndph. ox. ac. uk 即可。② 使用 uknconv 工具生成包含有数据下载链接的.bulk 文件，命令为：ukbconv ukb23456.enc_ukb bulk -s20207；其中 20207 表示数据的编号。我们将会得到一个 ukb23456.bulk 文件，需要注意的是，无论数据的编号怎么换，这个.bulk 文件的名字是不会变的，所以使用 ukbconv 下载新的编号前需要将此文件更名或者转存。③ 使用 ukbfetch 工具下载数据文件。大批量数据下载后是以研究对象为单位，一个研究对象为一个压缩包或一个文件。使用命令：./ukbfetch -ukb23456.bulk -s$CN -m1000；其中 $CN 表示下载开始的对象编号，m1000 表示的是从开始编号逐 1 递增的后 1 000 个对象数据。批量数据没法一次性将所有数据下载下来，需要一批一批的下载。下载后的 20207 编号数据为多个压缩包，解压后为多个.dcm 文件。

4.3 基因数据提取

基因数据的提取只能在 Linux 平台中进行，使用命令为：./gfetch 22828 -c1 -ak12345r23456.key，并且需要把.enc_ukb 后缀的数据包文件放在同一目录下。其中 22828 是一个基因数据的编号，c1 表示 1 号染色体。基因数据是根据指定染色体下载

表 2　UKB 数据下载的基本步骤

序号	步骤	备注
1	从项目管理员邮箱中获取32位MD5检验码和密钥文件	密钥为.key文件
2	从官方网站Projects-->View/edit-->data中下载数据包与6个基本工具	数据包为.enc后缀文件
3	使用ukbmd5工具检查数据包的完整性，命令：ukbmd5 ukb23456.enc	
4	使用ukbunpack工具解密数据包，命令：ukbunpack ukb23456.enc k56789r23456.key；	解密后数据包为.enc_ukb格式
5	使用ukbconv、ukbfetch、gfetch或ukblink下载相应的数据	不同的数据有不同的下载方式

的，一个编号的大小最大可达 200 G，根据具体的数据类型确定。比如单倍型 22418 全染色体总的占用空间为 91.5G，但是基因插补数据 22828 的 1 号染色体就有 181G，全染色体数据更是达到了 2T 的占用空间。总体来说，染色体编号越小，占用空间越大。详细的基因数据介绍见相关参考文献[20]。

4.4　其他数据提取

除了前述的几个主要类型数据，UKB 还包括了一些记录级的医院和初级保健数据——这可通过展示页面的下载页面中的数据门户进行访问。返回的数据集——来自研究者在研究中使用了 UKB 数据，但没有直接纳入主要资源，主要使用工具 ukblink 下载，下载方式可在官网检索资源 655 中找到。

5　UKB 数据库研究方向

为了解目前使用 UKB 数据库进行相关研究的现况，本研究围绕 UKB 数据库相关研究进行了可视化分析。以 Web of Science 为例，以“UK Biobank*”为关键词进行检索，排除不相关文献，最后将得到的所有文献进行整合，并将近 5 年的整体文献发表情况进行可视化分析。截至目前，UKB 数据库中主要的研究方向以基因遗传学、神经科学、心血管系统、计算机科学等方向为主，见图 3。

从研究热点的角度看，将满足出现频率大于 3 次的关键词以关系网络展示，见图 4。其中，每一个节点代表一个关键词，节点的圆圈直径越大则关键词出现的频次越高，不同关键词之间连接的线越粗，表明两者之间的关系越紧密。显然，从已发表的文献中可知，心血管疾病（cardiovascular disease）、新型冠状病毒性肺炎（COVID-19）、房颤（atrial fibrillation）、慢性肾脏疾病（chronic kidney disease）、精神性疾病（mental health）是目前聚焦的疾病病种，并且关联较为紧密的影响和调节因素中包括认知功能（cognition）、营养饮食（diet）、血压（blood pressure）、睡眠质量（sleep duration）、体育活动（exercise）、教育水平（education）等[21]。

从研究方法的角度，孟德尔随机化研究（Mendelian randomization）、全基因组关联研究（genome-wide association study），机器学习（machine learning）、队列研究（cohort study）等是目前 UKB 数据库研究的热点方法，孟德尔随机化研究与全基因组关联研究主要针对基因数据库的 GWAS 数据，确定和评估不同疾病的相关遗传学变异和位点，从而探索基因表达多态性的遗传变异机制；机器学习作为人工智能技术与医学图像结合的学习方法，在影像分割、影像分类及预测肿瘤的良恶性等方面发展迅速，有助于实现基于图像的个性化医疗决策；而队列研究是国际上公认的探讨常见重大疾病病因最有效的方法，也是研究遗传和其他暴露因素与健康结局的重要临床研究方法之一，尤其在 UKB 中的健康相关数据中应用最为广泛。

表 3　结构化数据提取样例

eid	100021-0.0	100021-1.0	100021-2.0	100021-3.0	100021-4.0
1000010		0.98		0.36	
1000076					
1000087*	0.79	1.54	9.86	5.14	
1000091		3.28	2.39		
1000104	18.67				
1000118	0.22			10.36	
1000120					
1000147		1.76	0.05	3.72	
1000162					0.02
1000171					
1000185	1.02			6.4	2.6

eid：患者编号；100021：维生素D的数据编码；*：观察eid为1000087的患者，可发现在4次访问评估中心中，都自我报告维生素D的指标。

由此，UKB 作为一项前瞻性群体研究计划，涉及了相当广泛而丰富的多学科问题，并且在神经系统疾病和心血管疾病研究领域具有巨大的潜力，但仍存在很多新的研究方向需要研究人员去探讨和思考。

6　人工智能与 UKB 数据结合

目前，围绕代谢组学、生物信息学、医学影像学、系统生物学等领域，结合人工智能的 UKB 数据库挖掘推动了精准医学的发展，因而将 UKB 和人工智能相关的主题词作为关键词，将所有文献中出

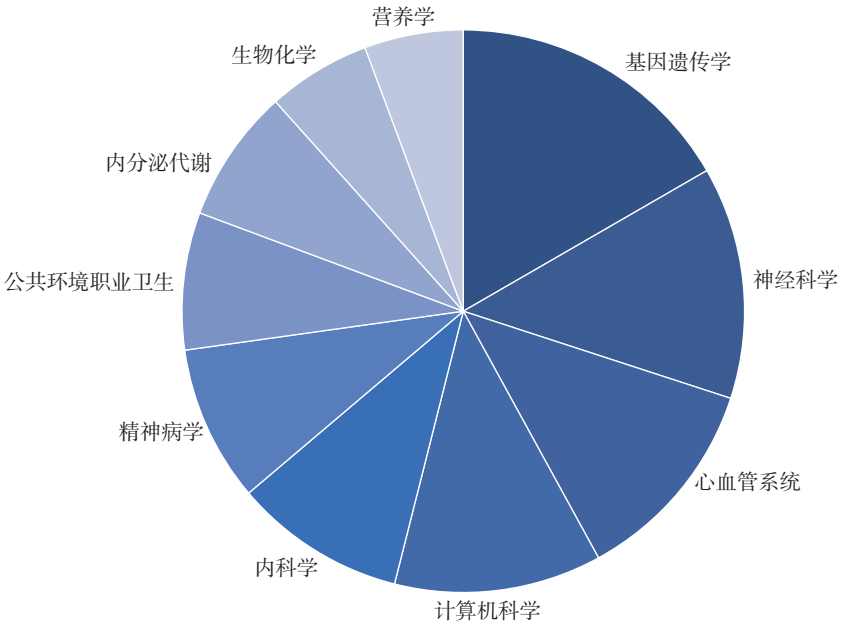

图 3　UKB 基于 Web of Science 研究方向分布图

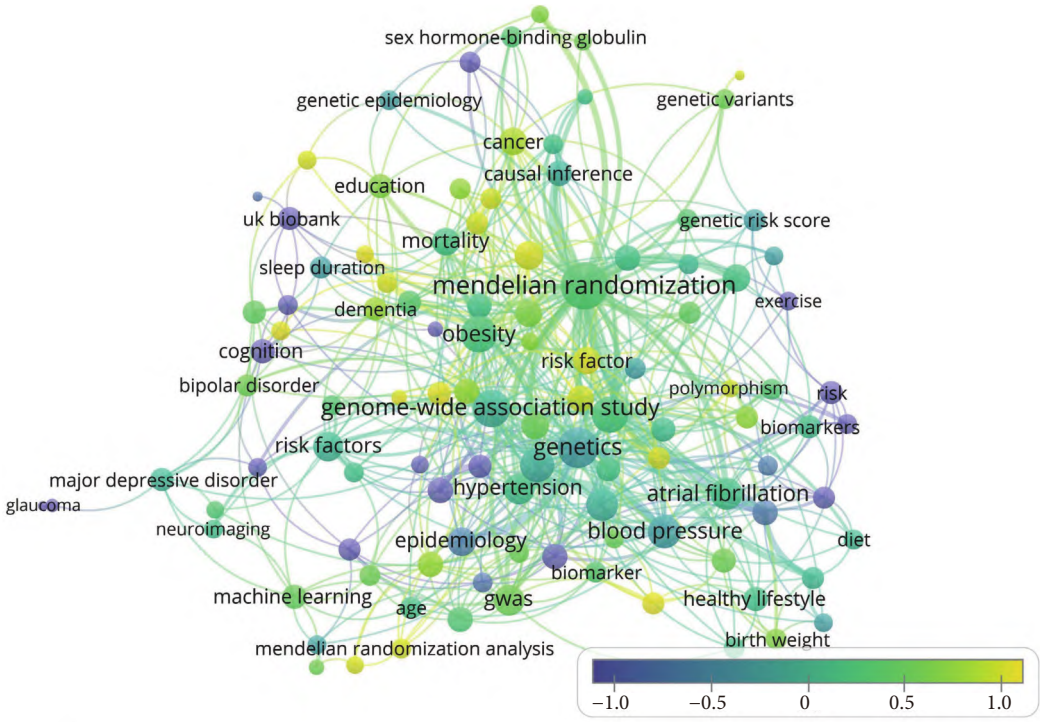

图 4　UKB 关键词分布图谱

现 3 次以上的关键词在可视化工具 VOSviewer 下转化为密度热图展示见（图 5）。我们发现，机器学习和深度学习是最主要的人工智能学习方法，图像分割、分类、预测模型等研究手段与 MRI、FMRI 联系紧密。另外，研究领域囊括了心血管疾病、脑梗死、阿尔茨海默症及抑郁症等热点问题。目前最新的研究成果揭示了 UKB 数据库未来的发展方向，例如机器学习被应用于心脏 MRI，识别主动脉瓣畸形和其他不良的心脏预后事件，同时进行了全自动左心室分析，试图评估自动化左心室的质量和体积[22]；使用机器学习方法对客观测量的睡眠和身体活动行为进行最大的评估，有助于了解治疗的有效性及与行为变异相关的疾病过程[23]；基于 UKB 的遗传数据建立心肌梗塞患病风险的预测模型[24]；应用最新的卷积神经网络模型结合 UKB 原始神经影像数据预测脑年龄、认知水平及老化程度等[25]。人工智能领域的技术，特别是深度学习方法作为最前沿的研究方向，囊括了计算机视觉、自然语言处理

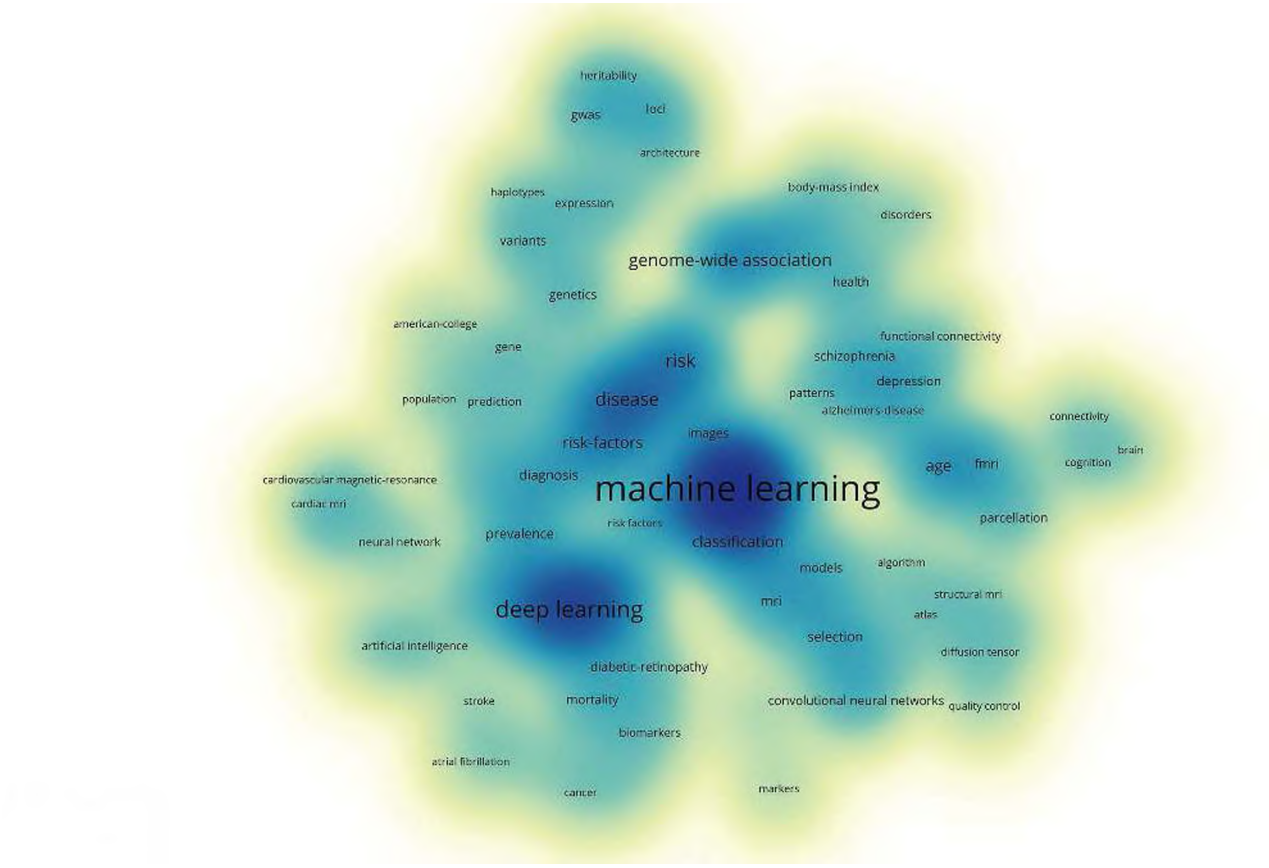

图 5　UKB 与人工智能相关的关键词热图

的研究手段，从计算机视觉的角度，不同于传统的计算机辅助检测，深度学习通过对目标图像特征的提取和识别，实现图像的分割、预测病灶的严重程度及定性定位；从自然语言处理的角度，不仅可通过 UKB 文本数据快速地进行语义分析，并且实现文本的转换提取，将临床电子数据和描述报告标签化[26]。因而，如何更深入地将影像数据与基因数据进行结合，探究大脑结构和功能、遗传突变、从遗传到影像病理机制的因果联系，是所有研究者们将不断探讨的重要研究方向。

7　讨论

UKB 是一项前瞻性的大规模以健康人群为基础的研究，旨在评估 40～69 岁正常人群中、老年疾病的所有遗传和非遗传因素，UKB 不仅提供了详细的健康信息，并且收集了大量的基线数据和样本，提供了血液、尿液样本、认知测试、在线问卷、健康记录、运动心电图活动监测、全基因组基因表型数据，尤其在 2013 年起收集了 10 万参与者的多模态成像数据，并将进行持续 20 年的随访记录。目前已经囊括 8 500 多人死亡、750 000 多例流行病和突发癌症等共计 60 多万人住院的相关数据，同时与一系列其他数据库建立联系。UKB 正在开发对一系列疾病领域的结果和遗传基因表型进行准确识别和亚分类，为影像基因组学研究提供了相关依据。UKB 数据库的优势在于，有海量的健康人群高通量基因组学数据和神经影像数据，完整的影像数据足足有数十 TB[3,17]，并且研究者无需对复杂的神经影像再进行分析处理。另外，UKB 中还包括了 COVID-19 患者数据，可结合神经系统疾病的预后和认知功能进行长期随访研究，将基因组学数据和影像数据进行整合[27-28]。

UKB 数据库作为一个大型的公共数据库，目前向所有的符合有关伦理和科学标准的研究人员开放。从近几年的研究来看，研究者们以基因表型数据和行为与临床结局的相关性为主要研究方向，研究主题涵盖了新型冠状病毒性肺炎、心血管疾病、神经退行性疾病、精神性疾病等疾病。随着影像数据库的扩展和逐渐成熟，基因组和影像数据的结合将是未来可能的研究方向，研究者将尝试探索从遗传突变到影像指标和神经或精神疾病的机制，深入了解正常和紊乱大脑功能和行为在遗传表型特征、遗传变异的表现，利用最新的人工智能技术，使大数据与深度学习融合在基因影像学领域获

得突破性进展。

本研究对 UKB 进行了整体而系统的介绍，从 UKB 的数据库内容、提取数据流程、具体研究应用和与人工智能结合的研究发展等方面探索，有助于更多研究人员系统了解 UKB 数据库，以期为心血管疾病、神经系统疾病及传染病等疾病的预防和治疗开辟新的研究领域。

参考文献

1 UK Biobank (2006). Protocol for a large-scale prospective epidemiological resource. Available at: www.ukbiobank.ac.uk/resources/.

2 Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics, 2005, 6(6): 639-646.

3 Littlejohns TJ, Holliday J, Gibson LM, *et al*. The UK Biobank imaging enhancement of 100, 000 participants: rationale, data collection, management and future directions. Nat Commun, 2020, 11(1): 2624.

4 Miller KL, Alfaro-Almagro F, Bangerter NK, *et al*. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci, 2016, 19(11): 1523-1536.

5 穆琳, 裴昀, 冬冬. 深度学习在骨关节医学影像中的研究及应用进展. 临床放射学杂志, 2020, 39(8): 1666-1669.

6 万艳丽, 雷行云, 王岩, 等. 基于层次化深度学习的海量医学影像组织与检索研究. 医学信息学杂志, 2015, 36(5): 46-51.

7 McIntosh AM, Stewart R, John A, *et al*. Data science for mental health: a UK perspective on a global challenge. Lancet Psychiatry, 2016, 3(10): 993-998.

8 Wilkinson T, Schnier C, Bush K, *et al*. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol, 2019, 34(6): 557-565.

9 The UK Biobank. Genotyping and quality control of UK Biobank, a Large-Scale, extensively phenotyped prospective resource. Available at: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf.

10 The UK Biobank. UK Biobank Axiom Array Content Summary. Available at: https://tools.thermofisher.cn/content/sfs/brochures/uk_axiom_biobank_contentsummary_brochure.pdf.

11 Bycroft C, Freeman C, Petkova D, *et al*. The UK Biobank resource with deep phenotyping and genomic data. Nature, 2018, 562(7726): 203-209.

12 Tyrrell J, Jones SE, Beaumont R, *et al*. Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. BMJ, 2016, 352: i582.

13 Van Hout CV, Tachmazidou I, Backman JD, *et al*. Exome sequencing and characterization of 49, 960 individuals in the UK Biobank. Nature, 2020, 586(7831): 749-756.

14 Sudlow C, Gallacher J, Allen N, *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med, 2015, 12(3): e1001779.

15 Hanscombe KB, Coleman JRI, Traylor M, *et al*. ukbtools: an R package to manage and query UK Biobank data. PLoS One, 2019, 14(5): e0214311.

16 Smith SM, Elliott LT, Alfaro-Almagro F, *et al*. Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. Elife, 2020, 9: e52677.

17 Cox SR, Lyall DM, Ritchie SJ, *et al*. Associations between vascular risk factors and brain MRI indices in UK Biobank. Eur Heart J, 2019, 40(28): 2290-2300.

18 林岚, 熊敏, 吴水才. 英国生物银行在神经影像领域应用的研究综述. 生物医学工程学杂志, 2021, 38(3): 594-601.

19 Fry A, Littlejohns TJ, Sudlow C, *et al*. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol, 2017, 186(9): 1026-1034.

20 Bycroft C, Freeman C, Petkova D, *et al*. Genome-wide genetic data on ~500, 000 UK Biobank participants. 2017. DOI: 10.1101/166298.

21 Daghlas I, Dashti HS, Lane J, *et al*. Sleep duration and myocardial infarction. J Am Coll Cardiol, 2019, 74(10): 1304-1314.

22 Suinesiaputra A, Sanghvi MM, Aung N, *et al*. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. Int J Cardiovasc Imaging, 2018, 34(2): 281-291.

23 Willetts M, Hollowell S, Aslett L, *et al*. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96, 220 UK Biobank participants. Sci Rep, 2018, 8(1): 7961.

24 Khera AV, Chaffin M, Aragam KG, *et al*. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. 2017. DOI: 10.1101/218388.

25 Buchanan CR, Bastin ME, Ritchie SJ, *et al*. The effect of network thresholding and weighting on structural brain networks in the UK Biobank. Neuroimage, 2020, 211: 116443.

26 李莉, 黄韬, 王新宇, 等. 胸腔X射线影像数据库——MIMIC-CXR数据探索. 中国循证心血管医学杂志, 2021, 13(6): 653-656,660.

27 Elliott LT, Sharp K, Alfaro-Almagro F, *et al*. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature, 2018, 562(7726): 210-216.

28 Hassan L, Peek N, Lovell K, *et al*. Disparities in COVID-19 infection, hospitalisation and death in people with schizophrenia, bipolar disorder, and major depressive disorder: a cohort study of the UK Biobank. Mol Psychiatry, 2022, 27(2): 1248-1255.

收稿日期：2022-04-29 修回日期：2022-07-15
本文编辑：张洋