arXiv:2302.00606v1 [cs.CL] 1 Feb 2023

RW3D：用于理解疫情心理影响的多模态面板数据集

Isabelle van der Vegt[1,3,†] **andBennett Kleinberg**[2,3,†]

[1]乌得勒支大学，社会学系，乌得勒支大学，荷兰乌得勒支 3584 CH[2]蒂尔堡大学，方法论与统计学系，荷兰蒂尔堡 5037 AB[3]伦敦大学学院，安全与犯罪科学系，英国伦敦 WC1E6BT[*]通讯作者：Isabelle van der Vegt (i.w.j.vandervegt@uu.nl)

[†]这些作者对本研究的贡献相同

摘要

除了对公共卫生造成深远影响外，新冠肺炎疫情也对世界各地的人们产生了显著的心理影响。为进一步了解这一问题，我们介绍现实世界担忧波动数据集（RW3D）。该数据集采用重复测量设计，汇集了英国三年间（2020年：$n = 2441$，2021年：$n = 1716$ 和2022年：$n = 1152$）收集的丰富开放式自由文本回答，以及有关情绪、重大生活事件和心理压力源的调查数据。本文介绍数据收集流程、记录的变量、参与者的人口统计学特征，以及从数据中提取的高阶心理变量和基于文本的派生变量。RW3D是一项独特的一手数据资源，有望启发关于疫情心理影响的新研究问题，尤其是那些探讨不同模态（此处指文本数据、心理调查变量和人口统计学特征）如何随时间关联的问题。

背景与概要

自疫情暴发以来，社会与行为科学家一直在收集数据，研究新冠肺炎及其相关措施对个体造成的心理影响。这场全球卫生危机严重影响了世界各地人们的生活。与此同时，它也为不同领域的社会科学家提供了契机，使他们得以研究人类面对前所未有的情境时会如何应对。由此产生了多篇论文及相关数据集，其中包括采用心理学视角的研究。例如，*COVIDiSTRESS GlobalSurvey* 收集了感知压力、对当局的信任以及对抗疫措施的遵从程度等指标；调查数据于2020年3月30日至5月30日期间，在39个国家和地区的173,426名个体中收集[1]。同样，*PsyCorona* 数据集包含疫情初期收集的数据（$n = 34,526$），涵盖全球41个社会，测量了心理变量和行为，例如离家外出和保持身体距离[2]。该数据集已被用于后续研究，例如测量合作与跨社会信任[3]，以及情绪与新冠肺炎风险感知之间的关联19[4]。另有研究探讨了“疫情疲劳”这一概念（即人们感到自己无法“坚持”遵守限制措施）；目前已有来自八个国家的数据[5]。研究发现，疫情疲劳与限制措施的严厉程度有关，而且疫情疲劳还会引发政治不满。

尤其值得关注的是自由文本回答，它们有助于我们了解个体在疫情期间及疫情之后的境况；与目标明确的问卷式数据收集相比，自由文本回答能够更深入、更全面地涵盖各类主题。一些项目利用并公开了有关疫情影响的语言数据，通常为 Twitter 数据[6,7]。另一项研究分析了 Reddit 数据和调查数据，以衡量疫情期间心理状态的变化[8]。然而，这两种数据都来自不同的参与者，因此无法通过关联调查与文本模态，深入探索文本作者的真实心理状态。出于多种原因，向同一批参与者收集文本和调查数据十分可取。首先，自由文本回答使参与者能够不受限制地报告自己在疫情中的经历，从而有可能更深入地洞察心理过程。其次，同期收集的调查回答能够为可能影响文本表达的心理变量提供真实情况的测量。第三，自然语言处理领域的进步使人们能够对文本数据开展深入的定量分析，因此文本数据不再只是通常由人工进行定性分析的资源，其用途也得以拓展。然而，收集能够将文本维度与调查数据关联起来的数据成本高昂，因为这需要开展原始数据收集，无法通过“现成数据”（例如社交媒体帖子）实现。因此，迄今为止，此类数据集十分稀缺，而这类数据的缺乏也阻碍了我们研究疫情的心理影响。

本文填补了这一空白，并介绍了现实世界担忧波次数据集（*RW3D*）。该数据集独特地结合了关于情绪的真实调查数据，以及描述与疫情相关情绪的自由文本回答。该数据集内容丰富，使我们能够考察例如情绪反应、应对方式，以及因新冠肺炎而产生的担忧内容。鉴于潜在研究问题范围广泛，且此类数据来源既稀缺又不可或缺，我们将该数据集向研究群体开放。下文将详细介绍数据收集流程、记录的变量、参与者人口统计特征，以及样本流失分析和描述性统计。本文旨在详述这一独特资源，为丰富多样的研究问题提供启发。

方法

伦理

数据收集已获伦敦大学学院院系伦理审查委员会批准。研究未收集参与者的个人数据，所有参与者均已提供知情同意，同意参与研究并将其数据用于研究目的。

程序

该数据集分三波收集[1]，分别于2020年、2021年和2022年4月进行。数据收集于2020年4月在众包平台 Prolific Academic 上启动，初始样本量为 $n = 2500$。一年后，我们通过该众包平台联系同一批参与者开展后续数据收集，并为第一波中收集到数据的所有参与者开放参与名额。又过一年后，我们对第二波中收集到数据的参与者重复了这一流程。

在所有数据收集阶段，参与者均获知研究的性质。参与者首先填写自评情绪问卷并进行单一情绪选择，随后完成文本表达，最后提供控制变量（第1波和第2波）、生活事件和心理压力源变量（仅第3波，见图 1）。

只有根据该网站预筛选题[2]使用 Twitter 的英国境内的 Prolific 用户才有资格参与。完成任务后，每位参与者获得 0.50 英镑的报酬。尽管实际完成任务所花费的时间比预期略长，但为了避免在重复测量设计中引入报酬变化这一混杂变量，我们没有调整报酬。该任务通过 Qualtrics 实施。

时间安排与社会背景

第一波数据收集于2020年4月初进行，当时英国正处于封锁状态，死亡人数不断上升。伊丽莎白二世女王刚刚向全国发表讲话，英国首相鲍里斯·约翰逊因新冠肺炎症状住院（见[9]）。在第2波（2021年4月）期间，英国许多人已经接种了疫苗，学校、零售业和酒店餐饮业开始（部分）重新开放。冠状病毒的德尔塔变异株当时刚刚被发现（见[10]）。最后，在第3波（2022年4月）期间，所有针对入境英国人员的旅行限制都已解除，奥密克戎变异株正在迅速传播，有关派对门事件的新闻仍在持续发酵（即围绕封锁期间在唐宁街举行聚会而引发的政治丑闻）[11]。

Wave	初始	Age	女性（%）	最终	留存率（%）
2020	2500	33.84	65.15	2441	100.00
2021	1839	36.22	67.37	1716	70.30
2022	1227	37.10	68.40	1152	67.13

表1。各波次的样本基本特征。

人口统计变量

我们从 Prolific 获取了参与者的人口统计学特征。这些数据由注册参与者自愿提供，包括年龄、性别、出生国家、国籍、第一语言、就业状况、学生身份、出生国家、居住国家，以及他们在众包平台上的参与情况（已完成并获批准的任务数量）。我们在调查问卷中新增了一道人口统计学问题，询问他们的母语（因为母语可能与第一语言不同）。

2022年4月，参与者的平均年龄为37.10岁（标准差 = 11.98），其中68.40%为女性（31.42%为男性，其余参与者选择了不愿透露），见表1。绝大多数参与者（90.54%）表示其出生国家为英国，并将英国列为其

[1]请注意，此处的波次是指研究设计中的波次，不应与新冠肺炎疫情的波次混淆。[2]此处指在“以下哪些社交媒体网站是您定期（至少每月一次）使用的？”这一问题中选择了 Twitter 的参与者

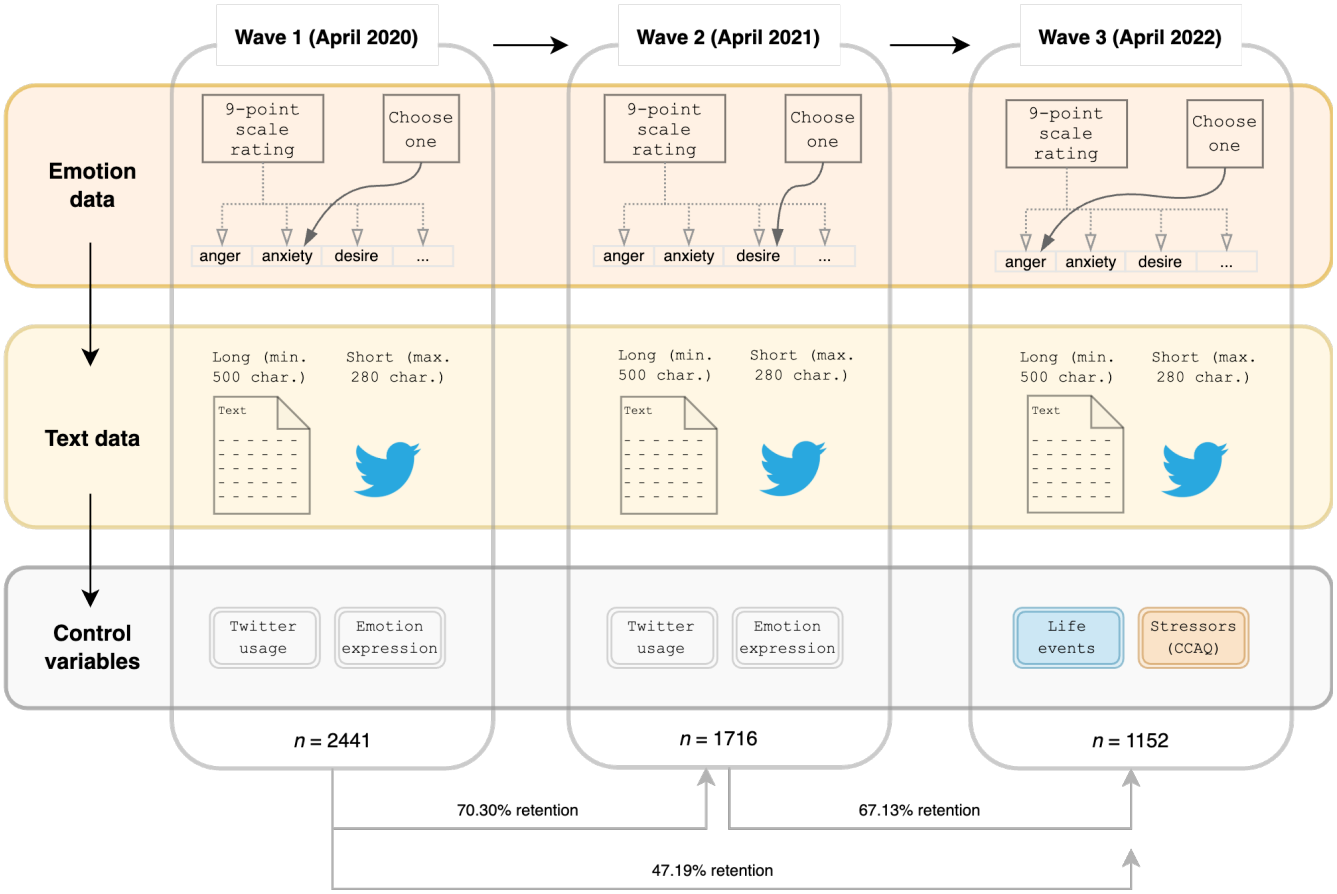

图 1. 三波数据收集的研究设计。参与者首先以 9 点量表对八种情绪进行评分，并选出最符合自己情况的一种，随后撰写一段长文本和一段短文本，之后填写有关控制变量的问卷（第1波和第2波），或提供其生活事件和心理压力源的详细信息（第3波）。人口统计变量来自 Prolific。

当前居住国（99.65%）。后者与我们进行的招募预筛选相符。关于就业状况，2020年，52.43%的人表示全职就业，22.74%从事兼职工作，10.50%没有有偿工作（例如已退休）。值得注意的是，2022年全职工作者的比例有所下降（42.35%）。同样，学生比例也从2020年的16.93%下降至2022年的10.94%。

情绪数据

自评情绪

参与者被要求使用9点评分量表，指出他们对新冠疫情状况有多担忧（1=完全不担忧； 5=中度担忧； 9=非常担忧），并说明此刻他们对新冠疫情状况的感受。对于后者，他们需要指出自己对以下八种情绪的感受有多强烈（1=完全没有； 5=中度； 9=非常强烈）：愤怒、厌恶、恐惧、焦虑、悲伤、幸福、放松、欲望[12]。评分通过滑块进行，每次以1为步长。

单一情绪选择

在上面列出的八种情绪中（即不包括担忧），每位参与者都被问到：“如果只能选择一种，以下哪种情绪最能代表你此刻的感受？”

表2 展示了情绪变量（自评量表值和离散选择）的描述性统计。总体趋势显示情况有所改善：积极情绪增加，消极情绪减少，但其中也存在一些隐藏的模式。先前的研究发现，参与者从2020年过渡到 2021[10] 的过程中形成了不同的聚类；下文将提供有关亚组的更多证据。

情绪	M_1	SD_1	比例.1	M_2	SD_1	比例.2	M_3	SD_3	*prop*.3
担忧	6.67	1.70	-	5.07	2.03	-	3.98	2.16	-
愤怒	3.76	2.18	0.04	3.47	2.35	0.08	2.85	2.19	0.06
厌恶	3.06	2.12	0.01	2.79	2.16	0.02	2.39	2.03	0.03
fear	5.63	2.30	0.09	3.77	2.30	0.02	2.85	2.05	0.02
焦虑	6.51	2.30	0.58	5.05	2.52	0.36	4.09	2.45	0.30
悲伤	5.55	2.31	0.15	4.64	2.57	0.19	3.48	2.35	0.13
幸福	3.55	1.84	0.01	4.29	1.98	0.05	4.76	2.10	0.07
放松	3.83	2.05	0.12	4.54	2.25	0.23	5.14	2.35	0.38
欲望	2.73	1.90	0.01	3.42	2.19	0.05	3.22	2.09	0.02

表2. 各波次自评情绪的描述性统计（M，SD）（量表： 1=完全没有； 5=中等程度； 9=非常强烈），以及选择相应情绪作为“最符合”情绪的个体比例。

文本数据

我们从每位参与者那里收集了两段文本回答。第一段文本数据通过以下指示获取：“请用几句话写下此时此刻你对新冠疫情状况的感受。这段文字应表达你此刻的感受。”参与者在文本框中输入回答；如果回答少于500个字符，系统会提示他们。随后，我们直接收集第二段文本回答，旨在获取一段较短、推文长度的文本，指示如下：“假设你必须用一条推文（最多280个字符）表达你当前对新冠疫情状况的感受。请在下方文本框中填写。”在这种情况下，如果参与者输入的文本少于10个字符或多于280个字符，系统会提示他们。

变量	M_1	SD_1	范围1	M_2	SD_2	范围2	M_3	SD_3	范围3
长：词元	126.17	39.37	[62, 1067]	125.17	32.05	[78, 387]	122.58	27.64	[62, 358]
长：字符数	624.20	196.66	[500, 5454]	621.27	158.19	[500, 2023]	610.82	133.98	[500, 1796]
简短：词元	26.63	15.90	[1, 75]	24.70	15.40	[1, 75]	24.59	14.62	[1, 67]
简短：字符数	131.11	76.97	[10, 281]	121.64	75.12	[10, 281]	122.01	73.36	[10, 282]

表 3. 语料库描述统计（均值、标准差、范围），涵盖数据收集的全部三个波次。请注意，字符数超过 280 是因为 rs 简短文本中的计算首尾空白字符时存在差异

语料库描述统计（表3）显示，在三个波次中，长文本和短文本的长度均保持稳定。语料库共包含430,751个词元（2020年：145,348；2021年：144,191；2022年：141,212）。下文展示两位参与者在三个波次中所写的文本示例（图2）。

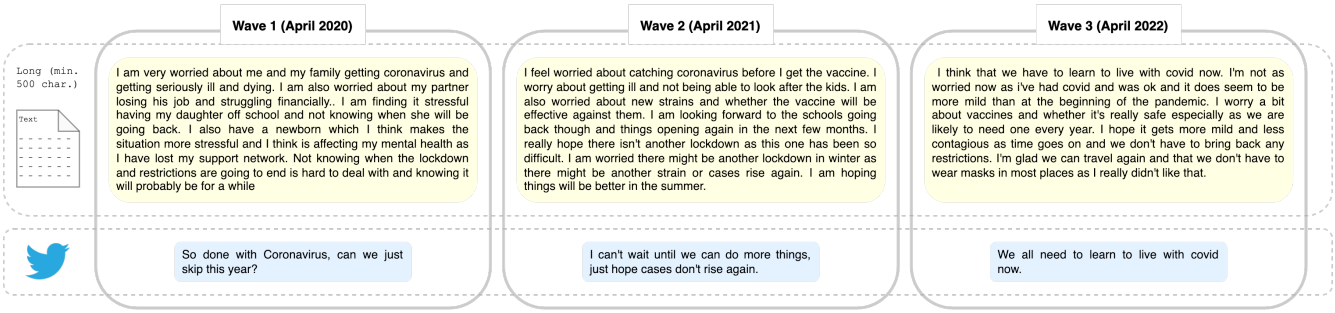

图2. 单个参与者的文本数据（较长文本和推文长度文本）

第1波和第2波的控制变量

在前两个波次（2020年4月和2021年），我们记录了两组控制变量：自我评定的文本情绪表达能力和Twitter使用情况。我们决定在第三波中不再收集这些变量。之所以舍弃这些变量，是因为我们认为它们在个体内部变化不大，而且我们已经有两次测量（第1波和第2波），且两次测量之间存在较强相关性（补充材料表1）。

情绪表达

作为调查问卷中自我报告的情绪与文本中的情绪表达之间关联的一项潜在控制变量，我们请参与者使用9点评分量表（1=完全不能； 5=中等程度； 9=非常好）评估自己：i）总体上表达感受的能力，ii）在推文长度文本中表达感受的能力，以及iii）在较长文本中表达感受的能力。

*Twitter*使用情况

作为专门针对推文长度文本的另一项潜在混杂变量，我们询问了参与者的Twitter使用情况。参与者使用9点评分量表（1=从不；5=每月； 9=每天）说明自己：i）使用Twitter的频率，ii）发布推文的频率，以及iii）在Twitter上参与对话的频率。

第3波新增变量

最近一次波次（2022年4月）增加了两个构念，取代了前几次波次中的控制变量。为了更好地了解疫情期间可能影响参与者情绪调适及其文本表达的调节变量，我们收集了疫情期间重要生活事件的数据，并使用了一份危机应对问卷[13]。

生活事件

我们回顾性地询问参与者，在过去两年里，他们的生活中是否发生过任何重要事件或变化。首先，我们询问他们：“在过去两年里，你的生活中是否发生过任何事情——无论是积极还是消极的——影响了你应对新冠疫情状况的方式？”[] [] [] 回答“是”的参与者随后被要求描述该事件，注明事件发生的月份和年份，并在从 -10（非常负面）到 +10 （非常正面）的量表上评估该事件的影响。如果还有其他事件，参与者还可以再提交一件（最多两件）。之后，作者对所有生活事件进行了定性编码，归纳出总体类别。例如，被解雇、换工作和大学毕业后找到第一份工作都归入“工作”类别；结婚、找到伴侣和分手则都归入“恋爱关系”类别（更多示例请参见补充材料表4）。

三分之一的参与者（33.85%）报告称，在数据收集期间经历了重大生活事件。最常见的生活事件类别是“死亡”（例如，家庭成员去世），且几乎都被评为负面生活事件（97.59%）。与工作有关的生活事件（例如，换工作）也很常见，大多数参与者（69.88%）将其评为具有正向强度（表4）。其他生活事件，如“心理健康”（例如，经历惊恐发作、获得心理健康诊断）和“财务”（例如，还清贷款、收入损失），则呈现出更为矛盾的模式，被评为正面和负面的比例大致相同。大多数生活事件发生在2021年12月（中位数）。各生活事件类别的示例见补充材料表4。

危机期间的压力源

为测量心理压力源，我们使用了《危机应对评估问卷》（CCAQ）[13]中的一部分。具体而言，我们从两个角度提出了若干问题：过去两年中，他们对一系列问题有多担忧，以及这些问题最终变得有多棘手。对于每个角度，参与者均使用9点评分量表（1=完全没有让我担忧/完全不棘手； 9=让我极度担忧/极其棘手）对以下12项担忧进行作答：他们自身的身体健康、心理健康和安全，亲近之人的身体健康、心理健康和安全，失业，以及未

事件	比例	强度均值（标准差）	强度中位数	负向强度比例	正向强度比例
无生活事件	66.15	NA	NA	NA	NA
死亡	7.20	-8.43 (2.83)	-10 [-10;6]	97.59	2.41
job	7.20	3.08 (6.3)	6 [-10;10]	28.92	69.88
恋爱关系	3.12	-0.33 (7.95)	-4 [-10;10]	52.78	47.22
家庭	3.04	-1.88 (7.48)	-6 [-10;10]	65.71	37.14
生殖	3.04	5.94 (5.97)	8 [-10;10]	17.14	82.86
健康	2.52	-5 (6.38)	-8 [-10;10]	75.86	24.14
move	2.34	5.85 (4.93)	8 [-8;10]	11.11	88.89
家庭健康	1.91	-7.91 (2.58)	-8 [-10;0]	95.45	0.00
心理健康	1.48	-1.18 (7.65)	-4 [-10;10]	52.94	47.06
教育	0.69	-2.75 (5.23)	-5 [-8;4]	62.5	37.5
财务	0.61	-0.29 (7.61)	-4 [-10;10]	57.14	42.86
生活方式	0.43	9.6 (0.89)	10 [8;10]	0.00	100.00
友谊	0.26	1.33 (9.87)	6 [-10;8]	33.33	66.67

表4。 第三波收集的生活事件数据汇总，包括强度（M、SD、中位数）以及报告正向和负向强度的参与者比例，强度范围为-10（非常负向）至 +10 （非常正向）

能有足够的钱维持生计、获取日常基本用品（食物等）、社会动荡、与家人分离，以及亲近的人施暴。

对 CCAQ 的回答显示，参与者最担忧亲人的人身安全和心理健康。这些压力源在现实中发生的程度表明，参与者自身的心理健康及其亲人的心理健康受到了影响（表5）。对于所有测量的担忧，实际问题得分从未超过担忧得分。也就是说，参与者对某个问题的担忧始终高于该问题实际造成的问题程度。我们发现，担忧—问题差异在不同担忧事项之间的分布并不均匀；下文将提供证据，说明参与者可分为两个亚组。

变量	M_{worry}	SD_{worry}	M_{actual}	SD_{actual}
自身人身安全	5.16	2.23	3.58	2.34
自身心理状态	5.69	2.40	4.92	2.63
自身安全	4.73	2.24	2.72	1.96
亲人的人身安全	6.85	1.94	4.33	2.45
亲人的心理健康	6.31	2.09	4.77	2.40
亲人的安全	6.22	2.20	3.57	2.36
失业	3.45	2.52	2.45	2.24
财务问题	5.06	2.59	3.88	2.63
获取基本生活用品	4.65	2.32	3.55	2.33
社会动荡	4.45	2.14	3.10	2.13
与家庭分离	5.13	2.63	4.13	2.65
亲近的人遭遇暴力	1.67	1.49	1.44	1.30

表5. 对心理压力源的担忧程度以及每种压力源实际造成的问题程度汇总（M，SD），评分范围为1–9（1=完全不让我担忧/完全没有问题； 9=让我极度担忧/问题极其严重）

数据记录

RW3D 数据集可在开放科学框架上获取，网址为https://doi.org/10.17605/OSF.IO/6S2JP。相应的存储库（https://osf.io/9b85r/）包含所有补充材料，以及一本详细介绍完整数据集的变量代码手册，其中说明了命名规范。

技术验证

本节介绍 i）为确保数据质量而采取的步骤，包括参与者排除标准；以及 ii）如何获得数据衍生数据。

数据保留与排除

每轮数据收集后，我们根据两项基于文本的标准排除参与者：如果长文本不是用英语撰写的（通过 *cld* R 软件包[14]判定），或标点符号词元占比超过 20%，则排除该参与者。这两项标准对于确保文本数据质量至关重要。第二波和第三波的保留率分别为 70.30% 和 67.13%。表格1 展示了三波的样本量和主要样本描述性统计。

数据衍生数据

我们从数据中获得了两类衍生数据：一类基于文本数据，另一类基于情绪和 CCAQ 问卷。通过文本数据，我们得到了高阶主题，为每篇书面文本提供总体主题，可用于研究参与者所写的内容。心理变量（情绪量表和 CCAQ）被映射到高阶心理构念，这些构念由参与者在情绪转变（从 2020年 到 2021年，以及从 2021年 到 2022年）以及担忧—问题差异方面形成的聚类所表征。

主题

为了捕捉文本数据中的总体主题，我们使用 *stm* R 软件包[15]，针对每个数据收集波次的文本数据构建了相关主题模型。该概率模型基于这样的假设：每段文本都由多个主题混合而成，而每个主题又由若干词语混合构成，每个词语属于某个主题的概率各不相同[15,16]。表6 展示了各波次长文本中出现频率最高的三个主题（完整的长、短文本主题和术语列表见补充材料表2和表3）。我们根据每个主题中出现频率最高的术语为各主题赋予了标签。

wave	主题	% 文档	术语
第1波	遵守规则	10.29	人们、感受、看到、许多人、死亡、政府、规则、认为、将会、遵守
第1波	这会持续多久?	10.16	将会、担忧、感受、熟悉、正常、恢复、漫长、知道、生活、希望
第1波	担心亲人	9.53	担忧、也、家庭、朋友、时间、焦虑、工作、家里、感觉、将会
第2波	期待恢复正常	15.04	将会、感觉、正常、希望、回来、获得、疫苗、事情、期待、向前
第2波	怀念正常生活	14.38	工作、想要、朋友、家人、感觉、见面、想念、时间、能够、家
第2波	焦虑	11.91	将会、担心、疫苗、忧虑、也、感觉、新冠肺炎、焦虑、影响、病毒
第3波	正常生活	9.99	正常、回归、现在、生活、活着、新冠肺炎、担心、感觉、得到、可以
第3波	仍然担忧	8.70	新冠肺炎、感觉、仍然、担心、得到、人们、检测、尽管、影响、不
第3波	新变异株	8.67	仍然，变异株、疫苗、将、病例、病毒、新、担忧、数量、人们

表6。排名前三的　每个波次长文本中最常见的主题，以及对应的标签和出现频率最高的词语。

高阶心理聚类

先前的研究发现，从第1波到第 2[10]波的情绪转变中，数据内部存在潜在聚类。我们评估了这个扩展数据集中是否还存在其他情绪聚类，以及疫情压力源引发的担忧与实际问题之间是否存在差异。对于每个概念，我们按照以下步骤进行（图3）：我们计算了两个转变时期的情绪评分变化值（第2波减去第1波，以及第3波减去第2波），并计算了 CCAQ 担忧评分与问题评分之间的差值（担忧减去实际问题）。随后，我们针对情绪转变1（2020年到2021年）、情绪转变2（2021年到2022年）以及担忧—问题差异，进行了 K均值聚类[17]。我们通过碎石图和轮廓系数法的趋同结果来确定聚类数量。在所有三个转变时期中，都发现了两个聚类。

2020—2021年过渡期的情绪聚类（表 7）呈现出两组特征：一组参与者（占样本的44.36%）的情绪健康显著改善；另一组（55.64%）的情绪反应则类似于听天由命（即这些参与者报告了更强烈的愤怒、厌恶和悲伤，但担忧和恐惧有所减轻）。随后一年的聚类再次显示出一组应对良好的参与者（40.45%），其模式与此前的良好应对组非常相似，唯一不同的是欲望没有增加。与之相对的是一组适应不良的参与者（59.55%），他们与听天由命聚类有一定重叠，但由于这一组的恐惧和焦虑增加、欲望减少，我们将其另行命名。

关于担忧与问题差异聚类（表8），规模较大的聚类（58.16%）表现出明显更强的“过度担忧”（即他们表示，对各种担忧事项的担忧程度远高于这些事项实际成为问题的程度）。

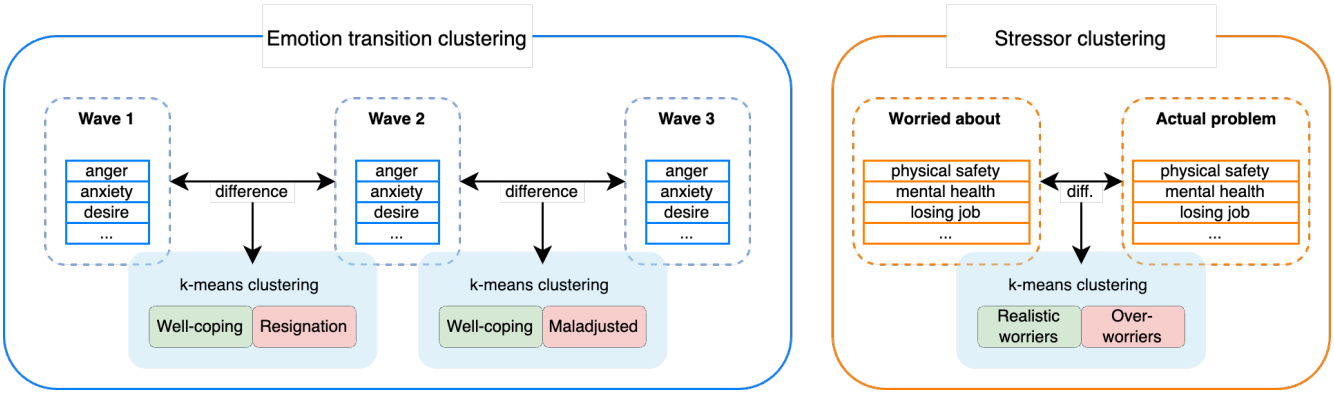

图3。 针对两种情绪转变和担忧—问题差异的聚类方法。根据肘部法和轮廓系数法，k均值算法显示了理想的聚类数 $k = 2$。

情绪	2021年良好应对	2021年顺从	2022年良好应对	2022年适应不良
愤怒	-1.68	0.83	-2.38	0.56
焦虑	-3.16	-0.11^{ns}	-2.82	0.30
欲望	1.24	0.25	0.27^{ns}	-0.52
厌恶	-1.26	0.53	-1.74	0.51
fear	-3.53	-0.53	-2.52	0.17
幸福	1.85	-0.14^{ns}	1.65	-0.34
放松	2.21	-0.48	2.24	-0.53
悲伤	-2.77	0.56	-3.27	0.28
担忧	-2.40	-0.97	-1.79	-0.62
Size	44.36%	55.64%	40.45%	59.55%

表7. 2020—2021年和2021—2022年转变期间各情绪转变聚类的情绪均值（除标注为*ns*者外，均与0有显著差异，$p < .01$，双侧检验）。两个应对良好的聚类呈现相似模式，表明积极情绪增加、消极情绪减少。听天由命（2021年）聚类与适应不良聚类有一定重叠，但在欲望和恐惧的变化方面有所不同。

这些过度担忧者在涉及亲人的身体健康和安全的问题上尤为明显。相比之下，我们称为现实型担忧者的群体（41.84%）始终表现出较低的担忧—问题差异。只有在涉及亲近的人施暴（家庭暴力）的问题上，两个群体的看法一致（担忧程度低，实际担忧也低）。

变量	现实型担忧者	过度担忧者
自身人身安全	0.76	2.74
自身心理健康	0.35	1.37
自身安全	1.17	3.17
亲人的身体健康	1.21	4.33
亲人的心理健康	0.69	2.72
亲人的安全	1.35	4.47
失业	0.59	1.57
财务问题	0.66	1.91
满足基本需求	0.53	1.88
社会动荡	0.85	2.04
与家人分离	0.54	1.64
家庭暴力	0.22	0.24
Size	41.84%	58.16%

表8。 担忧—问题差异的聚类（所有结果均与 0 显著不同，$p < .01$，双侧 ）

使用说明

理解并应对COVID-19疫情带来的心理影响，并可能为未来全球危机的影响做好准备，仍然是一项持续的研究挑战。其中一个障碍是缺乏高质量数据，无法将个体经历疫情的不同模态联系起来。当前数据集论文介绍了RW3D，这是一个针对英国参与者的重复测量数据集，将通过调查方法考察的心理变量与丰富的文本回答结合起来。人们对疫情期间应对方式的解释性关系仍缺乏充分了解。借助RW3D，我们可以通过面板模型考察生活事件、文本数据中提出的担忧，或社会人口特征的变化（例如失业）和变量（例如性别），在多大程度上能够解释个体转变为各种应对方式。深入了解这些复杂关系，也有助于我们更有针对性地向最需要的人提供干预措施。值得注意的是，数据集中纳入了各种控制变量，使研究人员能够控制潜在的混杂因素。

此外，我们还可以了解文本数据与心理变量之间关系的一些基本方面。通过将不同模态联系起来，我们可以检验能否根据文本数据预测真实情绪，以及滞后设计能否根据前几年的文本数据，帮助我们提前预判之后某个时点的情绪变化。同样，由于我们掌握了参与者的生活事件和压力源，因此可以评估这些信息能否从文本数据中提取出来。许多应用型基于文本的研究都隐含地假设，这些心理变量可以从文本数据中显现出来；然而，能够支持这种映射的丰富数据集十分稀缺。

局限性

使用此数据集时，需要考虑一些局限性。首先，数据仅收集自英国参与者。虽然这使得研究者能够结合英国特有的情况（例如感染传播、政府应对措施）对英国进行深入分析，但研究结果未必能推广至其他人群。其次，部分变量是回顾性收集的，要求参与者在重大生活事件发生最多两年后报告这些事件，并回想自己对危机应对的担忧和实际问题。第三，尽管人口统计学特征的分布相当广泛，但该数据集并未采用全国代表性样本。事后缓解这一问题的一种方法，或许是根据样本特征在英国人口中的占比进行加权（参见[18]）。

代码可用性

此数据描述文档没有关联的自定义代码。为进行数据（预）处理并获取数据衍生数据，我们使用了现有的 R 软件包，包括 *cld* 用于英语语言检查[14]，*quanteda*[19] 和 *stringr*[20] 用于文本元数据（字符数、词元数、标点符号）的处理，*stm* 软件包[15] 用于构建主题模型，以及 *factoextra*软件包[21] 用于确定聚类数量，以获得高阶心理构念。

参考文献

1. Yamada, Y. 等. COVIDiSTRESS全球调查数据集：COVID-19疫情暴发的心理和行为后果数据集。《科学数据》8, 3, 10.1038/s41597-020-00784-9（2021年）。编号：1，出版商：Nature出版集团。2. Kreienkamp, J., Agostini, M., Krause, J., Leander, a. N. P. & Collaboration, o. b. o. t. P. PsyCorona：对COVID-19的全球反应。《APS观察》33（2020年）。3. Romano, A. 等. COVID-19疫情期间不同社会间的合作与信任。《跨文化心理学杂志》52, 622–642, 10.1177/0022022120988913（2021年）。出版商：SAGE出版公司。4. Han, Q. 等. 疫情期间COVID-19风险感知与情绪及心理健康的关联。《情感障碍杂志》284, 247–255, 10.1016/j.jad.2021.01.049（2021年）。5. Jørgensen, F., Bor, A., Rasmussen, M. S., Lindholt, M. F. & Petersen, M. B. 疫情疲劳加剧了COVID-19疫情期间的政治不满。《美国国家科学院院刊》119, e2201266119（2022年）。6. Banda, J. M. 等. 用于开放科学研究的大规模COVID-19推特讨论数据集——一项国际协作。《流行病学》2, 315–324, 10.3390/epidemiologia2030024（2021年）。编号：3，出版商：多学科数字出版研究所。7. Naseem, U., Razzak, I., Khushi, M., Eklund, P. W. & Kim, J. COVIDSenti：用于COVID-19情感分析的大规模基准推特数据集。《IEEE计算社会系统汇刊》8, 1003–1015, 10.1109/TCSS.2021.3051189（2021年）。会议名称：《IEEE计算社会系统汇刊》。8. Ashokkumar, A. & Pennebaker, J. 社交媒体对话揭示了COVID-19暴发之初美国城市中发生的重大心理变化，10.1126/sciadv.abg7843（2021年）。

9. Kleinberg, B., van der Vegt, I. & Mozes, M. 测量COVID-19现实世界担忧数据集中的情绪。载于：ACL 2020首届 COVID-19 自然语言处理研讨会论文集（2020年）。10. Mozes, M., van der Vegt, I. & Kleinberg, B. 一项关于疫情一年后情绪反应的重复测量研究。《科学报告》11, 1–11（2021年）。11. Strick, K. 派对门——新冠疫情期间唐宁街派对丑闻时间线 | 《标准晚报》（2022年）。12. Harmon-Jones, C., Bastian, B. & Harmon-Jones, E. 离散情绪问卷：一种测量状态性自我报告情绪的新工具。《公共科学图书馆·综合》11, e0159915（2016年）。13. Lahlou, S. 等. CCAQ：一份共享的知识共享危机应对评估问卷。世界大流行病研究网络（2016年）。14. Ooms, J. cld3：谷歌的紧凑型语言检测器 3。15. Roberts, M. E., Stewart, B. M. & Tingley, D. stm：用于结构主题模型的 R 软件包。《统计软件杂志》91, 1–40, 10.18637/jss.v091.i02（2019年）。16. Blei, D. M. & Lafferty, J. D. 动态主题模型。载于：第23届国际机器学习会议论文集，113–120（2006年）。17. Likas, A., Vlassis, N. & Verbeek, J. J. 全局 k 均值聚类算法。《模式识别》36, 451–461（2003年）。18. Bradley, V. C. 等. 缺乏代表性的大型调查显著高估了美国疫苗接种率。《自然》600, 695–700（2021年）。19. Benoit, K. 等. quanteda：用于文本数据定量分析的 R 软件包。《开放源代码软件杂志》3, 774, 10.21105/joss.00774（2018年）。20. Wickham, H. stringr：用于常见字符串操作的简洁、一致的封装函数（2022年）。21. Kassambara, A. & Mundt, F. factoextra：提取并可视化多变量数据分析结果（2020年）。R 软件包版本 1.0.7。