

Time to treat the climate and nature crisis as one indivisible global health emergency

Chris Zielinski, on behalf of the authorship group listed below

To cite: Zielinski C. Time to treat the climate and nature crisis as one indivisible global health emergency. *BMJ Health Care Inform* 2023;**30**:e100938. doi:10.1136/bmjhci-2023-100938

Received 13 October 2023
Accepted 13 October 2023

Over 200 health journals call on the United Nations (UN), political leaders and health professionals to recognise that climate change and biodiversity loss are one indivisible crisis and must be tackled together to preserve health and avoid catastrophe. This overall environmental crisis is now so severe as to be a global health emergency.

The world is currently responding to the climate crisis and the nature crisis as if they were separate challenges. This is a dangerous mistake. The 28th Conference of the Parties (COP) on climate change is about to be held in Dubai while the 16th COP on biodiversity is due to be held in Turkey in 2024. The research communities that provide the evidence for the two COPs are unfortunately largely separate, but they were brought together for a workshop in 2020 when they concluded that: ‘Only by considering climate and biodiversity as parts of the same complex problem...can solutions be developed that avoid maladaptation and maximize the beneficial outcomes’.¹

As the health world has recognised with the development of the concept of planetary health, the natural world is made up of one overall interdependent system. Damage to one subsystem can create feedback that damages another—for example, drought, wildfires, floods and the other effects of rising global temperatures destroy plant life and lead to soil erosion, and so inhibit carbon storage, which means more global warming.² Climate change is set to overtake deforestation and other land-use change as the primary driver of nature loss.³

Nature has a remarkable power to restore. For example, deforested land can revert to forest through natural regeneration, and marine phytoplankton, which act as natural carbon stores, turn over one billion tonnes of photosynthesising biomass every 8 days.⁴ Indigenous land and sea management has a

particularly important role to play in regeneration and continuing care.⁵

Restoring one subsystem can help another—for example, replenishing soil could help remove greenhouse gases from the atmosphere on a vast scale.⁶ But actions that may benefit one subsystem can harm another—for example, planting forests with one type of tree can remove carbon dioxide from the air but can damage the biodiversity that is fundamental to healthy ecosystems.⁷

THE IMPACTS ON HEALTH

Human health is damaged directly by both the climate crisis, as the journals have described in previous editorials,^{8,9} and by the nature crisis.¹⁰ This indivisible planetary crisis will have major effects on health as a result of the disruption of social and economic systems—shortages of land, shelter, food and water, exacerbating poverty, which in turn will lead to mass migration and conflict. Rising temperatures, extreme weather events, air pollution and the spread of infectious diseases are some of the major health threats exacerbated by climate change.¹¹ “Without nature, we have nothing” was UN Secretary-General António Guterres’s blunt summary at the biodiversity COP in Montreal last year.¹² Even if we could keep global warming below an increase of 1.5°C over preindustrial levels, we could still cause catastrophic harm to health by destroying nature.

Access to clean water is fundamental to human health, and yet pollution has damaged water quality, causing a rise in waterborne diseases.¹³ Contamination of water on land can also have far-reaching effects on distant ecosystems when that water runs off into the ocean.¹⁴ Good nutrition is underpinned by diversity in the variety of foods, but there has been a striking loss of genetic diversity in the food system. Globally, about a fifth of people



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ.

Centre for Global Health,
University of Winchester,
Winchester, UK

Correspondence to

Dr Chris Zielinski, UK Health Alliance on Climate Change, London, WC1H 9JR, UK; chris.zielinski@ukhealthalliance.org

rely on wild species for food and their livelihoods.¹⁵ Declines in wildlife are a major challenge for these populations, particularly in low-income and middle-income countries. Fish provide more than half of dietary protein in many African, South Asian and small island nations, but ocean acidification has reduced the quality and quantity of seafood.¹⁶

Changes in land use have forced tens of thousands of species into closer contact, increasing the exchange of pathogens and the emergence of new diseases and pandemics.¹⁷ People losing contact with the natural environment and the declining biodiversity have both been linked to increases in non-communicable, autoimmune and inflammatory diseases, and metabolic, allergic and neuropsychiatric disorders.^{10 18} For Indigenous people, caring for and connecting with nature is especially important for their health.¹⁹ Nature has also been an important source of medicines, and thus reduced diversity also constrains the discovery of new medicines.

Communities are healthier if they have access to high-quality green spaces that help filter air pollution, reduce air and ground temperatures, and provide opportunities for physical activity.²⁰ Connection with nature reduces stress, loneliness and depression while promoting social interaction.²¹ These benefits are threatened by the continuing rise in urbanisation.²²

Finally, the health impacts of climate change and biodiversity loss will be experienced unequally between and within countries, with the most vulnerable communities often bearing the highest burden.¹⁰ Linked to this, inequality is also arguably fuelling these environmental crises. Environmental challenges and social/health inequities are challenges that share drivers and there are potential co-benefits of addressing them.¹⁰

A GLOBAL HEALTH EMERGENCY

In December 2022 the biodiversity COP agreed on the effective conservation and management of at least 30% of the world's land, coastal areas and oceans by 2030.²³ Industrialised countries agreed to mobilise \$30 billion per year to support developing nations to do so.²³ These agreements echo promises made at climate COPs.

Yet many commitments made at COPs have not been met. This has allowed ecosystems to be pushed further to the brink, greatly increasing the risk of arriving at 'tipping points', abrupt breakdowns in the functioning of nature.^{2 24} If these events were to occur, the impacts on health would be globally catastrophic.

This risk, combined with the severe impacts on health already occurring, means that the WHO should declare the indivisible climate and nature crisis as a global health emergency. The three preconditions for the WHO to declare a situation to be a public health emergency of international concern²⁵ are that it (1) is serious, sudden, unusual or unexpected; (2) carries implications for public health beyond the affected State's national border; and (3) may require immediate international action. Climate change would appear to

fulfil all of these conditions. While the accelerating climate change and loss of biodiversity are not sudden or unexpected, they are certainly serious and unusual. Hence we call for the WHO to make this declaration before or at the 77th World Health Assembly in May 2024.

Tackling this emergency requires the COP processes to be harmonised. As a first step, the respective conventions must push for better integration of national climate plans with biodiversity equivalents.³ As the 2020 workshop that brought climate and nature scientists together concluded, 'Critical leverage points include exploring alternative visions of good quality of life, rethinking consumption and waste, shifting values related to the human-nature relationship, reducing inequalities, and promoting education and learning'.¹ All of these would benefit health.

Health professionals must be powerful advocates for both restoring biodiversity and tackling climate change for the good of health. Political leaders must recognise both the severe threats to health from the planetary crisis as well as the benefits that can flow to health from tackling the crisis.²⁶ But first, we must recognise this crisis for what it is: a global health emergency.

List of Authors

Kamran Abbasi, Editor-in-Chief, *BMJ*; Parveen Ali, Editor-in-Chief, *International Nursing Review*; Virginia Barbour, Editor-in-Chief, *Medical Journal of Australia*; Thomas Benfield, Editor-in-Chief, *Danish Medical Journal*; Kirsten Bibbins-Domingo, Editor-in-Chief, *JAMA*; Stephen Hancocks, Editor-in-Chief, *British Dental Journal*; Richard Horton, Editor-in-Chief, *The Lancet*; Laurie Laybourn-Langton, University of Exeter; Robert Mash, Editor-in-Chief, *African Journal of Primary Health Care & Family Medicine*; Peush Sahni, Editor-in-Chief, *National Medical Journal of India*; Wadeia Mohammad Sharief, Editor-in-Chief, *Dubai Medical Journal*; Paul Yonga, Editor-in-Chief, *East African Medical Journal*; Chris Zielinski, University of Winchester.

This Comment is being published simultaneously in multiple journals. For the full list of journals see: <https://www.bmj.com/content/full-list-authors-and-signatories-climate-nature-emergency-editorial-october-2023>

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

REFERENCES

- Otto-Portner H, Scholes B, Agard J, *et al*. Scientific outcome of the IPBES-IPCC co-sponsored workshop on biodiversity and climate change. 2021.

- 2 Ripple WJ, Wolf C, Lenton TM, *et al.* Many risky feedback loops amplify the need for climate action. *One Earth* 2023;6:86–91.
- 3 European Academies Science Advisory Council. Key messages from European science academies for UNFCCC COP26 and CBD COP15. 2021. Available: <https://easac.eu/publications/details/key-messages-from-european-science-academies-for-unfccc-cop26-and-cbd-cop15> [Accessed 01 Oct 2023].
- 4 Falkowski P. Ocean science: the power of Plankton. *Nature* 2012;483:S17–20.
- 5 Dawson NM, Coolsaet B, Sterling EJ, *et al.* The role of indigenous peoples and local communities in effective and equitable conservation. *E&S* 2021;26.
- 6 Bossio DA, Cook-Patton SC, Ellis PW, *et al.* The role of soil carbon in natural climate solutions. *Nat Sustain* 2020;3:391–8.
- 7 Levia DF, Creed IF, Hannah DM, *et al.* Homogenization of the terrestrial water cycle. *Nat Geosci* 2020;13:656–8.
- 8 Atwoli L, Baqui AH, Benfield T, *et al.* Call for emergency action to limit global temperature increases, restore biodiversity, and protect health. *BMJ* 2021;374:n1734.
- 9 Atwoli L, Erhabor GE, Gbakima AA, *et al.* COP27 climate change conference: urgent action needed for Africa and the world. *BMJ* 2022;379:o2459.
- 10 WHO, UNEP, Convention on Biological D. Connecting global priorities: biodiversity and human health: a state of knowledge review. 2015. Available: <https://www.cbd.int/health/SOK-biodiversity-en.pdf> [Accessed 1 Oct 2023].
- 11 Magnano San Lio R, Favara G, Maugeri A, *et al.* How antimicrobial resistance is linked to climate change: an overview of two intertwined global challenges. *Int J Environ Res Public Health* 2023;20:1681.
- 12 Jelskov U. "Without nature, we have nothing": UN chief sounds alarm at key UN Biodiversity event [UN News]. 2022. Available: <https://news.un.org/en/story/2022/12/1131422> [Accessed 01 Oct 2023].
- 13 World Health Organization. State of the world's drinking water: an urgent call to action to accelerate progress on ensuring safe drinking water for all [World Health Organization]. 2022. Available: <https://www.who.int/publications/i/item/9789240060807> [Accessed 01 Oct 2023].
- 14 Comerós-Raynal MT, Brodie J, Bainbridge Z, *et al.* Catchment to sea connection: impacts of terrestrial run-off on Benthic Ecosystems in American Samoa. *Mar Pollut Bull* 2021;169:112530.
- 15 IPBES. Assessment report on the sustainable use of wild species. 2022. Available: <https://www.ipbes.net/sustainable-use-assessment>
- 16 Falkenberg LJ, Bellerby RGJ, Connell SD, *et al.* Ocean acidification and human health. *Int J Environ Res Public Health* 2020;17:4563.
- 17 Dunne D. Climate change "already" raising risk of virus spread between mammals. 2022. Available: <https://www.carbonbrief.org/climate-change-already-raising-risk-of-virus-spread-between-mammals/> [Accessed 01 Oct 2023].
- 18 Altıveş S, Yıldız HK, Vural HC. Interaction of the microbiota with the human body in health and diseases. *Biosci Microbiota Food Health* 2020;39:23–32.
- 19 Schultz R, Cairney S. Caring for country and the health of aboriginal and Torres Strait Islander Australians. *Med J Aust* 2017;207:8–10.
- 20 Macguire F, Mulcahy E, Rossington B. The lancet countdown on health and climate change - policy brief for the UK. 2022. Available: https://s41874.pcdn.co/wp-content/uploads/Lancet-Countdown-2022-UK-Policy-Brief_EN.pdf [Accessed 01 Oct 2023].
- 21 Wong FY, Yang L, Yuen JWM, *et al.* Assessing quality of life using WHOQOL-BREF: a cross-sectional study on the association between quality of life and neighborhood environmental satisfaction, and the mediating effect of health-related behaviors. *BMC Public Health* 2018;18:1113.
- 22 Simkin RD, Seto KC, McDonald RI, *et al.* Biodiversity impacts and conservation implications of urban land expansion projected to 2050. *Proc Natl Acad Sci U S A* 2022;119:e2117297119.
- 23 Secretariat of the Convention on Biological Diversity. COP15: nations adopt four goals, 23 targets for 2030 in landmark UN Biodiversity agreement [Convention on Biological Diversity]. 2022. Available: <https://www.cbd.int/article/cop15-cbd-press-release-final-19dec2022> [Accessed 01 Oct 2023].
- 24 Armstrong McKay DI, Staal A, Abrams JF, *et al.* Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science* 2022;377:eabn7950.
- 25 WHO guidance for the use of annex 2 of the International health regulations [World Health Organization]. 2005. Available: [https://www.who.int/publications/m/item/who-guidance-for-the-use-of-annex-2-of-the-international-health-regulations-\(2005\)](https://www.who.int/publications/m/item/who-guidance-for-the-use-of-annex-2-of-the-international-health-regulations-(2005)) [Accessed 01 Oct 2023].
- 26 Australian Government Department of Health and Aged Care. Consultation on Australia's first national health and climate strategy [Australian Government Department of Health and Aged Care]. 2023. Available: <https://www.health.gov.au/news/consultation-on-australias-first-national-health-and-climate-strategy> [Accessed 01 Oct 2023].

© 2023 Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ. <https://creativecommons.org/licenses/by/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Integrating digital health technologies into complex clinical systems

Mark Sujan  ^{1,2}

To cite: Sujan M. Integrating digital health technologies into complex clinical systems. *BMJ Health Care Inform* 2023;**30**:e100885. doi:10.1136/bmjhci-2023-100885

Received 30 August 2023
Accepted 26 September 2023

Modern health systems must embrace digital technologies to address challenges like ongoing shortages in the global health and care workforce, significant diagnostic backlogs and the requirements of diverse and ageing populations. The COVID-19 pandemic and the exceptional advances in artificial intelligence (AI) and machine learning (ML) have accelerated the drive towards digitalisation of health systems.¹ However, making digital health technologies work in practice remains challenging in terms of how these technologies are designed, how their performance and safety in operation are assured and how their impact on staff and on patients is assessed.²

A key problem undermining the successful implementation of digital health technology is the persistent focus on technology in isolation, which is at odds with the realities of complex health and care systems. The shortcomings of this technology-centric focus can be seen, for example, when reviewing the lack of successful clinical deployment of the multitude of ML algorithms developed during the pandemic to support the diagnosis and management of COVID-19.³ More broadly, the apparent success of ML algorithms found in retrospective evaluation studies is frequently not replicated in subsequent prospective studies.^{4,5} The difficulty of translating successful retrospective evaluation of algorithms into useful clinical practice has been referred to as the challenge of the last mile.⁶

Arguably, consideration of the challenge of the last mile, that is, of the realities of complex clinical systems, cannot be left to the end, but needs to inform the design of AI and, more generally, digital health technologies from the outset. The design of digital health technologies needs to be based on a systems perspective. A systems perspective considers how technology fits into the wider clinical system, where success depends on interactions with people, other

IT systems, the physical environment and the organisation of clinical and administrative processes.^{7,8} The two 'editor's choice' articles illustrate the importance of considering the sociotechnical nature of digital health technology implementation.

Hong and colleagues studied an ML tool to identify at-risk patients who are undergoing outpatient cancer treatment in order to reduce their acute care needs.⁹ The ML tool had been developed and implemented as part of a randomised controlled quality improvement project. The authors describe using a survey instrument on completion of the implementation phase to elicit perceptions from healthcare staff about the impact of the adoption of the ML tool and to identify practical implementation challenges. While, generally, feedback about the ML tool was positive and encouraging, the results highlight that the introduction of ML into a complex clinical system might affect different stakeholders in different and unevenly distributed ways. The need for prospective and mixed methods evaluation of algorithms has been recognised in the literature, but to date, there are few documented examples.¹⁰ Through their findings, Hong and colleagues demonstrate the importance of such empirical studies of AI in complex clinical settings.

In the second article, Richter and Ammenwerth¹¹ aim to support practitioners with the implementation of risk management for networked medical devices in hospitals based on the international standard IEC 80001. While the principles of risk management have been long established and documented in several standards, these principles are often expressed in abstract and conceptual terms. This leaves practitioners facing a challenging implementation gap.¹² Richter and Ammenwerth attempt to bridge this gap by providing a catalogue of 49 specific steps and things that practitioners must put in place, along with 18 indicators to assess the impact of risk management activities. This practical



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Investigation Education, Health Services Safety Investigation Body, Poole, UK

²Human Factors Everywhere, Woking, UK

Correspondence to

Dr Mark Sujan;
mark.sujan@gmail.com



guidance has been developed through a consensus exercise with experts and practitioners. The findings were then validated in a case study in one Austrian hospital, where parts of the catalogue were implemented and evaluated for effectiveness, complexity and satisfaction based on practitioner feedback. This approach can serve as an illustration and a blueprint for making best practice guidance practically relevant and meaningful in complex clinical environments.

Successful integration of digital health technologies into complex clinical systems requires a move away from a narrow and limiting technology focus towards a systems perspective, which needs to be reflected in the design, operation and evaluation of the technology. Practitioners need to be provided with meaningful tools and guidance to enable them to manage and to assess the operation of digital health technologies, and to ask the right questions of developers. The recent British Standard BS 30440, which outlines an auditable validation framework for healthcare AI, is another example of this.¹³ Finally, we need to continue efforts to build capacity and capability within health and care organisations to enhance their readiness to deploy such technologies meaningfully, for example, in the case of the National Health Service in England by extending training opportunities with NHS England (the former NHS Digital team) on digital clinical safety and AI safety or with the Health Services Safety Investigation Body on system-based investigation methods. Other health and care systems should develop similar education and training frameworks and opportunities.

Twitter Mark Sujan @MarkSujan

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD




Mark Sujan <http://orcid.org/0000-0001-6895-946X>

REFERENCES

- 1 Peek N, Sujan M, Scott P. Digital health and care in pandemic times: impact of COVID-19. *BMJ Health Care Inform* 2020;27:e100166.
- 2 Sujan M, Scott P, Cresswell K. Digital health and patient safety: technology is not a magic wand. *Health Informatics J* 2020;26:2295–9.
- 3 Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest Radiographs and CT scans. *Nat Mach Intell* 2021;3:199–217.
- 4 Nagendran M, Chen Y, Lovejoy CA, *et al*. n.d. Artificial intelligence versus Clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*:m689.
- 5 Wong A, Otlés E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065–70.
- 6 Coiera E. The last mile: where artificial intelligence meets reality. *J Med Internet Res* 2019;21:e16323.
- 7 Sujan M, Furniss D, Grundy K, *et al*. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* 2019;26:e100081.
- 8 Sujan M, Pool R, Salmon P. Eight human factors and Ergonomics principles for Healthcare artificial intelligence. *BMJ Health Care Inform* 2022;29:e100516.
- 9 Hong JC, Patel P, Eclow NCW, *et al*. Healthcare provider evaluation of machine learning-directed care: reactions to deployment on a randomised controlled study. *BMJ Health Care Inform* 2023;30:e100674.
- 10 Wu E, Wu K, Daneshjou R, *et al*. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021;27:582–4.
- 11 Richter S, Ammenwerth E. IT risk management for medical devices in hospital IT networks: a catalogue of measures and indicators. *BMJ Health Care Inform* 2023;30:e100639.
- 12 Habli I, White S, Sujan M, *et al*. What is the safety case for health IT? A study of assurance practices in England. *Safety Science* 2018;110:324–35.
- 13 Sujan M, Smith-Frazer C, Malamateniou C, *et al*. Validation framework for the use of AI in Healthcare: overview of the new British standard BS30440. *BMJ Health Care Inform* 2023;30:e100749.

© 2023 Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Digital health and care: emerging from pandemic times

Niels Peek ,^{1,2} Mark Sujan ,³ Philip Scott ⁴

To cite: Peek N, Sujan M, Scott P. Digital health and care: emerging from pandemic times. *BMJ Health Care Inform* 2023;**30**:e100861. doi:10.1136/bmjhci-2023-100861

Received 23 July 2023
Accepted 20 September 2023

ABSTRACT

In 2020, we published an editorial about the massive disruption of health and care services caused by the COVID-19 pandemic and the rapid changes in digital service delivery, artificial intelligence and data sharing that were taking place at the time. Now, 3 years later, we describe how these developments have progressed since, reflect on lessons learnt and consider key challenges and opportunities ahead by reviewing significant developments reported in the literature. As before, the three key areas we consider are digital transformation of services, realising the potential of artificial intelligence and wise data sharing to facilitate learning health systems. We conclude that the field of digital health has rapidly matured during the pandemic, but there are still major sociotechnical, evaluation and trust challenges in the development and deployment of new digital services.

INTRODUCTION

It is often blithely noted that the pandemic accelerated the uptake of digital capabilities that had unnecessarily languished in pilot status for many years, almost as though the smashing of cultural and organisational inertia was a ‘silver lining’ of the pandemic cloud. However, cautions and challenges remain to be considered, and we should not regard technology as a ‘silver bullet’ that can magic away the fundamental and long-standing issues in global healthcare. Our review takes a primarily UK focus, but we believe that many of the principles have wider application. Also, while we focus on the National Health Service (NHS), it is pertinent to note that the entire health and care sector stands to benefit from meaningful digital transformation.

FURTHER DIGITAL TRANSFORMATION IS NEEDED TO MAKE THE NHS FUTURE-PROOF

The COVID-19 pandemic catalysed rapid adoption of digital technology in the NHS¹ and resulted in significant changes to service delivery, primarily to enable remote working and reduce the risk of infection transmission but also to free up capacity in acute hospitals.² Primary care in particular saw a huge increase in remote consultations. There was also a surge in patients’ uptake of the NHS App, NHS

login and e-prescription services. Initially, these changes were positively perceived by the public, equating these changes with progress and improved efficiency and safety, in a service that was overdue for modernisation. As the pandemic progressed, however, there were growing concerns that remote consultations could lead to missed diagnoses, create challenges to therapeutic relationships and exacerbate health inequalities.³

In a recent review of 63 studies on primary care online consultation systems (11 of which were conducted during the pandemic), there was no quantitative evidence for the negative impact of online consultations on patient safety, but qualitative studies suggested varied perceptions of their safety.⁴ Online consultations increased access to care and decreased patient costs but were also sometimes found to have negative impacts on provider costs, staff and patient workloads, patient satisfaction and care equity. For instance, some primary care staff have indicated that they believe that patients seek help more readily via online consultations than they would have done via office-based consultations, and this leads to increasing staff workload.

Also, several remote monitoring models were widely implemented during the pandemic, such as COVID Oximetry @home⁵ and COVID virtual wards.⁶ Remote monitoring models ask patients to record health readings at home while these readings are reviewed and responded to by professionals elsewhere. There is typically an increased responsibility for patients to self-manage, for example, in the COVID Oximetry @home programme patients were expected to escalate care if their oxygen level dropped below certain thresholds.⁷ Large-scale evaluations of these programmes are still ongoing, but a rapid mixed-methods study found that many patients required support and preferred human contact, especially for identifying problems.⁸

Going forward a key challenge for the NHS is to clear the backlog of elective care that already existed before the COVID-19



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Centre for Health Informatics, The University of Manchester, Manchester, UK

²NIHR Applied Research Collaboration Greater Manchester, The University of Manchester, Manchester, UK

³Human Factors Everywhere Ltd, Woking, UK

⁴Institute of Management and Health, University of Wales Trinity Saint David, Swansea, UK

Correspondence to

Dr Niels Peek;
niels.peek@manchester.ac.uk

pandemic but was strongly exacerbated by it. For instance, an independent review of diagnostic services, commissioned by NHS England in 2019, revealed that diagnostic capacity (in terms of equipment and workforce) was much lower in England than in other developed countries.⁹ This is now hampering recovery from the pandemic. A major programme of work is underway to improve access to a wide range of diagnostic tests, with the establishment of community diagnostic centres being a key component of this programme.¹⁰ These centres have the potential to move routine diagnostic services closer to patients and reduce unnecessary hospital visits, but they risk exacerbating the existing workforce crisis in the NHS. It is, therefore, important to consider the wider sociotechnical system, allowing workforce investment to be focused in those places that can have the most impact and will, in turn, improve job satisfaction and retention.¹¹ The NHS also aims to improve the efficiency of follow-up in outpatient care. Long waiting times, delayed appointments and rushed consultations had already become common before the pandemic, but the number of patients waiting for a first appointment with a specialist is now more than seven million.¹² NHS England has set the ambition that 5% of outpatient attendances will be moved to patient-initiated follow-up pathways by March 2023—a target that is likely to increase in the future.¹³ Patient-initiated follow-up pathways allow patients to initiate outpatient follow-up appointments on an ‘as required’ basis compared with the traditional ‘physician-initiated’ model. Evidence on these pathways is still scarce but there are indications that they result in fewer overall outpatient appointments while maintaining equivalent if not better patient satisfaction, quality of life and clinical outcomes.¹⁴ There is ample opportunity to integrate artificial intelligence (AI) tools into these pathways, but this is an area that still needs development. We elaborate on this topic in the next section.

MAKING AI WORK IN PRACTICE REQUIRES A SYSTEMS APPROACH

The pandemic has surfaced structural and cultural problems that persist with the development and deployment of AI and machine learning (ML) in healthcare more widely. Healthcare is a complex sociotechnical system, and the current data and technology-centric focus needs to be complemented by a systems perspective. A systems perspective considers from the outset the impact of integrating AI tools into the wider clinical system, where interactions with people, other information systems, the physical environment and the organisation of clinical and administrative processes will be determinants of success.^{15 16}

During the pandemic, we saw an explosion in the number of ML algorithms to support the diagnosis and treatment of COVID-19. Examples include the use of Deep Learning to develop algorithms for the identification of COVID-19 from chest X-rays and CT scans,¹⁷ for

the identification of patients at risk of critical COVID-19-related disease progression¹⁸ and for the rapid triage of patients with COVID-19.¹⁹ However, in retrospect, there were few, if any, examples of successful clinical deployments of these algorithms.²⁰ Hence, we need to be cautious with the claims being made.

The tremendous push towards the development of AI during COVID-19 likely had several drivers, including the urgency to deal with the impact of COVID-19 as well as the collective research focus of the worldwide community, including funding sources, on COVID-19. But arguably, another key driver was sheer data availability. As the number of people infected with COVID-19 continued to grow, so did the number of data points that could be used to train algorithms. This was facilitated further by national efforts, such as the national COVID-19 chest imaging database established in the UK by NHSX.²¹ Developers can access this national database for performance and fairness testing of algorithms on a dataset representative of the UK population. While in principle, the availability of such national datasets is helpful to reduce the risk of bias of algorithms and to assess their performance, we need to be mindful that we do not create situations where developers simply train algorithms based on datasets that happen to be available rather than based on the need for and intended use of their models. The starting point for the development of algorithms should be an identified clinical need and an understanding of the associated clinical system to ensure that algorithms address clinically meaningful purposes. Then, suitable and high-quality data can be procured.

As the field of healthcare AI matures, we have seen welcome developments around reporting guidelines for ML algorithms, such as STARD-AI²² and PROBAST-AI²³ for reporting on the development and testing of diagnostic and prognostic prediction models, and SPIRIT-AI²⁴ and CONSORT-AI²⁵ for clinical trials of healthcare AI technologies. An important gap was addressed recently with the DECIDE-AI guideline,²⁶ which addresses early-stage, small-scale clinical evaluation of ML algorithms. The apparent lack of successful clinical deployment of the multitude of COVID-19 algorithms is a case in point—we cannot assume that retrospective evaluation of ML algorithms translates smoothly into successful adoption and deployment in clinical systems. We require a suitable empirical evaluation of AI and ML tools, which considers how these tools are integrated and used in specific clinical contexts. Developers can draw on recent guidance, such as the British Standard BS 30440, which formulates a comprehensive auditable validation framework for healthcare AI.²⁷

SHARING DATA WISELY BUILDS TRUST AND SUPPORTS LEARNING HEALTH SYSTEMS

Emergency expansion of data sharing was a pivotal part of the pandemic response, crucial to the unparalleled collaborative open science that made such

remarkable and rapid progress. Deidentified data linkage at the national level by programmes such as CVD-COVID-UK^{28 29} has enabled truly population-based analysis in ways that had previously been imagined but seldom realised. Data analytics has contributed to policy decisions, operational efficiencies and public health outcomes. Achieving this required innovative legislation, appropriate information governance and capable data infrastructure.

In Taiwan, for example, post-SARS legislation in 2007 introduced powers for governmental access to personal data in the event of emerging infectious diseases.³⁰ This empowered a task force to analyse diverse sources including COVID-19 test results, mobile device geolocation and hospital respiratory illness diagnosis tracking to provide remarkably powerful contact tracing and surveillance. Other countries that had rapid success deriving important insights from national data sharing during the worst of the pandemic were Scotland, Iceland, Israel and Qatar.³¹

However, data alone do not save lives.³² The best exemplars of data sharing are in fact forms of learning health system, where virtuous cycles comprising ‘practice to data’, ‘data to knowledge’ and ‘knowledge to practice’ have operated.³³ All this has required coherent policy support and adequate infrastructure, in terms of connectivity, storage, analytics and workforce. The countries that were most successful were able to build on existing foundations. The lesson here is that public health requires an ‘always-on’ infrastructure, ready to support the next inevitable pandemic.

‘Big data’ in health and care continues to have serious data quality issues, necessitating extensive cleansing, and often translation between heterogeneous data structures and coding schemes.³⁴ In many health systems, even fundamentals like patient matching between disparate data sets remain problematic.³⁵ Some health services, such as primary care in England, have financial incentive schemes that motivate standardised recording and coding³⁶ but despite this, the practice of clinical coding remains highly variable.³⁷ This poor data quality is one aspect of the problem of being ‘data rich, but information poor.’³⁸

How has the public reacted to more ‘open’ use of their health data? This seems to relate to how actually ‘open’ the data re-use is perceived to be, in the sense of transparency about who has access to what, under what rules, in what form and for what purpose. In the UK, a major data science corporation that was awarded significant NHS contracts in the pandemic is still regarded as ‘contentious’,³⁹ no doubt partly due to its involvement in past scandals about racist profiling in US law enforcement algorithms.⁴⁰ A proposed national extraction of data from general practice patient records in England, repeatedly delayed for various reasons, generated serious concerns from professional bodies due to its poor engagement with citizens about risks and benefits. On the other hand, citizens’ juries have proved to be a powerful method to enable genuine dialogue with the public and obtain specific measures of relative trust in a range of data-sharing initiatives.⁴¹

PREPARING FOR THE NEXT PANDEMIC

We suggest the following strategies to improve preparedness for future pandemics. First of all, further integration of telehealth services and remote patient monitoring technologies would enable seamless continuation of services with minimal physical contact during the next pandemic. We have only just started on this journey. In particular, a thorough evaluation of these services on care processes and patient outcomes is still needed. Second, a robust and interconnected data infrastructure, enabling real-time collection, analysis and sharing of health data across NHS and social care providers would facilitate early detection of outbreaks, rapid response coordination and effective resource allocation. The NHS is making progress on this front through the Secure Data Environment programme,⁴² but there is still a long way to go. Third, learning from mistakes during the COVID-19 pandemic, AI researchers should form multidisciplinary collaborations with provider organisations, social scientists and applied health researchers to define meaningful scenarios where ML algorithms can have added benefit during a pandemic, and develop methods and tools to address those scenarios when the time comes. Fourth and finally, building sufficient trust from both the public and the care professions is essential to become a truly data-driven and knowledge-driven learning health system that is prepared for the next pandemic.

Twitter Mark Suján @MarkSujan

Contributors All authors contributed equally to conceptualisation, writing and reviewing.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Niels Peek <http://orcid.org/0000-0002-6393-9969>

Mark Suján <http://orcid.org/0000-0001-6895-946X>

Philip Scott <http://orcid.org/0000-0002-6289-4260>

REFERENCES

- 1 Peek N, Suján M, Scott P. Digital health and care in pandemic times: impact of COVID-19. *BMJ Health Care Inform* 2020;27:e100166.
- 2 Hutchings R. *The impact of Covid-19 on the use of digital technology in the NHS*. London: Nuffield Trust, 2020.
- 3 Mroz G, Papoutsis C, Rushforth A, et al. Changing media Depictions of remote consulting in COVID-19: analysis of UK newspapers. *Br J Gen Pract* 2021;71:e1–9.
- 4 Darley S, Coulson T, Peek N, et al. Understanding how the design and implementation of online consultations affect primary care quality: systematic review of evidence with recommendations

- for designers, providers, and researchers. *J Med Internet Res* 2022;24:e37436.
- 5 National Health Service. COVID-19 guidance NOTE: COVID Oximetry @ home; 2022.
 - 6 National Health Service. Standard operating procedure: COVID virtual ward; 2022.
 - 7 Vindrola-Padros C, Singh KE, Sidhu MS, *et al*. Remote home monitoring (virtual wards) for confirmed or suspected COVID-19 patients: A rapid systematic review. *EClinicalMedicine* 2021;37:100965.
 - 8 Walton H, Vindrola-Padros C, Crellin NE, *et al*. Patients' experiences of, and engagement with, remote home monitoring services for COVID-19 patients: A rapid Mixed-Methods study. *Health Expect* 2022;25:2386–404.
 - 9 Richards M. *Diagnostics: Recovery and Renewal - Report of the Independent Review of Diagnostic Services for NHS England*. London: NHS England, 2020.
 - 10 Richards M, Maskell G, Halliday K, *et al*. Diagnostics: a major priority for the NHS. *Future Healthc J* 2022;9:133–7.
 - 11 Combes J, Crumpton E, Sujan M. Building better care the Ergonomist: chartered Institute of Ergonomics and human factors; 2022. 30–1.
 - 12 QualityWatch. NHS performance Tracker; 2022. Nuffield trust and the health foundation
 - 13 Reed S, Crellin N. *Patient-initiated follow-up: will it free up capacity in outpatient care?* Nuffield Trust, 2022.
 - 14 Taneja A, Su'a B, Hill AG. Efficacy of Patient-Initiated Follow-Up clinics in secondary care: a systematic review. *Intern Med J* 2014;44:1156–60.
 - 15 Sujan M, Furniss D, Grundy K, *et al*. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* 2019;26:e100081.
 - 16 Sujan M, Pool R, Salmon P. Eight human factors and Ergonomics principles for Healthcare artificial intelligence. *BMJ Health Care Inform* 2022;29:e100516.
 - 17 Li L, Qin L, Xu Z, *et al*. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* 2020;200905.
 - 18 Ustebay S, Sarmis A, Kaya GK, *et al*. A comparison of machine learning Algorithms in predicting COVID-19 Prognostics. *Intern Emerg Med* 2023;18:229–39.
 - 19 Liang W, Yao J, Chen A, *et al*. Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* 2020;11:3543.
 - 20 Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest Radiographs and CT scans. *Nat Mach Intell* 2021;3:199–217.
 - 21 Jacob J, Alexander D, Baillie JK, *et al*. Using imaging to combat a pandemic: rationale for developing the UK national COVID-19 chest imaging database. *Eur Respir J* 2020;56:2001809.
 - 22 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.
 - 23 Collins GS, Dhiman P, Andaur Navarro CL, *et al*. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and Prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.
 - 24 Rivera SC, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370:m3210.
 - 25 Liu X, Rivera SC, Moher D, *et al*. n.d. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ*:m3164.
 - 26 Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904.
 - 27 Sujan M, Smith-Frazer C, Malamateniou C, *et al*. Validation framework for the use of AI in Healthcare: overview of the new British standard Bs30440. *BMJ Health & Care Informatics* 2023;30:e100749.
 - 28 Wood A, Denholm R, Hollings S, *et al*. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021;373:n826.
 - 29 Ross JS. Covid-19, open science, and the CVD-COVID-UK initiative. *BMJ* 2021;373:898.
 - 30 Chen C-M, Jyan H-W, Chien S-C, *et al*. Containing COVID-19 among 627,386 persons in contact with the diamond princess cruise ship passengers who Disembarked in Taiwan: big data Analytics. *J Med Internet Res* 2020;22:e19540.
 - 31 Dhami S, Thompson D, El Akoum M, *et al*. Data-enabled responses to Pandemics: policy lessons from COVID-19. *Nat Med* 2022;28:2243–6.
 - 32 Scott P, Emerson K, Henderson-Reay T. Data saves lives. *BMJ* 2021;374:1694.
 - 33 Friedman CP. What is unique about learning health systems *Learn Health Syst* 2022;6:e10328.
 - 34 Scott PJ, Dunscombe R, Evans D, *et al*. Learning health systems need to bridge The 'Two cultures' of clinical Informatics and data science. *BMJ Health Care Inform* 2018;25:126–31.
 - 35 Guardiolle V, Bazoge A, Morin E, *et al*. Linking BIOMEDICAL data warehouse records with the National mortality database in France: large-scale matching algorithm. *JMIR Med Inform* 2022;10:e36711.
 - 36 NHS Digital. Quality and outcomes framework; 2022.
 - 37 Martin PM, Sbaffi L. Electronic health record and problem lists in Leeds, United kingdom: variability of general practitioners' views. *Health Informatics J* 2020;26:1898–911.
 - 38 Bergerum C, Petersson C, Thor J, *et al*. We are data rich but information poor': how do patient-reported measures stimulate patient involvement in quality improvement interventions in Swedish hospital departments. *BMJ Open Qual* 2022;11:e001850.
 - 39 Carding N. Contentious' US Tech firm to harvest patient data in NHSE waiting list push. *HSJ* 2022.
 - 40 Crawford K. Atlas of AI. In: *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 6 April 2021.
 - 41 NIHR applied research collaboration greater Manchester. In: *Citizens' Juries on Health Data Sharing in a Pandemic*. 2021.
 - 42 Department of Health and Social care. Secure data environment for NHS health and social care data - policy guidelines; 2022.

© 2023 Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards

Richard HR Roberts ^{1,2,3}, Stephen R Ali,^{1,3} Hayley A Hutchings,² Thomas D Dobbs,^{1,3} Iain S Whitaker^{1,3}

To cite: Roberts RHR, Ali SR, Hutchings HA, *et al*. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. *BMJ Health Care Inform* 2023;**30**:e100830. doi:10.1136/bmjhci-2023-100830

Received 14 June 2023

Accepted 05 September 2023

ABSTRACT

Introduction Amid clinicians' challenges in staying updated with medical research, artificial intelligence (AI) tools like the large language model (LLM) ChatGPT could automate appraisal of research quality, saving time and reducing bias. This study compares the proficiency of ChatGPT3 against human evaluation in scoring abstracts to determine its potential as a tool for evidence synthesis.

Methods We compared ChatGPT's scoring of implant dentistry abstracts with human evaluators using the Consolidated Standards of Reporting Trials for Abstracts reporting standards checklist, yielding an overall compliance score (OCS). Bland-Altman analysis assessed agreement between human and AI-generated OCS percentages. Additional error analysis included mean difference of OCS subscores, Welch's t-test and Pearson's correlation coefficient.

Results Bland-Altman analysis showed a mean difference of 4.92% (95% CI 0.62%, 0.37%) in OCS between human evaluation and ChatGPT. Error analysis displayed small mean differences in most domains, with the highest in 'conclusion' (0.764 (95% CI 0.186, 0.280)) and the lowest in 'blinding' (0.034 (95% CI 0.818, 0.895)). The strongest correlations between were in 'harms' ($r=0.32$, $p<0.001$) and 'trial registration' ($r=0.34$, $p=0.002$), whereas the weakest were in 'intervention' ($r=0.02$, $p<0.001$) and 'objective' ($r=0.06$, $p<0.001$).

Conclusion LLMs like ChatGPT can help automate appraisal of medical literature, aiding in the identification of accurately reported research. Possible applications of ChatGPT include integration within medical databases for abstract evaluation. Current limitations include the token limit, restricting its usage to abstracts. As AI technology advances, future versions like GPT4 could offer more reliable, comprehensive evaluations, enhancing the identification of high-quality research and potentially improving patient outcomes.

INTRODUCTION

In the dynamic landscape of medical research, clinicians face the daunting challenge of staying abreast of the latest advancements amid their demanding clinical responsibilities. The rate and varying quality of emerging research further compounds this challenge. A

number of appraisal tools exist to help readers assess the quality of the reported research, although these can also be time-consuming to employ and are at risk of user bias. The use of large language models (LLMs) like ChatGPT has the potential to automate this evaluation, thereby aiding clinicians in making informed decisions.¹ However, the accuracy of LLMs compared with human expertise as a gold standard remains uncertain. In November 2023, OpenAI unveiled ChatGPT, a generative pretrained transformer (GPT) language model grounded in transformer architecture, which empowers it to process vast amounts of text data and generate coherent text outputs by discerning the relationships between input and output sequences. ChatGPT has been trained on extensive human language datasets, and several studies attest to its ability to produce high-quality, coherent text outputs.²⁻³ Clinical research applications of ChatGPT have yielded promising results, suggesting that artificial intelligence could potentially critically appraise abstracts and liberate valuable clinician time.⁴ The objective of this study is to compare the proficiency of ChatGPT3, the third iteration of OpenAI's GPT model, in scoring abstracts against human evaluation as the benchmark. By determining the accuracy and efficiency of these LLMs in assessing research quality, we aim to explore their potential as valuable tools for clinicians in appraisal and evidence synthesis.

METHODS

In this study, we used a previously published paper as the basis of our comparison with ChatGPT.⁵ In their study, abstracts from a systematic review on implant dentistry were scored using the Consolidated Standards of



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ.

¹Reconstructive Surgery and Regenerative Medicine Research Centre, Swansea University, Swansea, UK

²Swansea University Medical School, Swansea University, Swansea, UK

³Welsh Centre for Burns and Plastic Surgery, Morriston Hospital, Swansea, UK

Correspondence to

Dr Richard HR Roberts; 838272@swansea.ac.uk

A The OCS is a measure of how many of the CONSORT-A items below are included in a given abstract. Each item below is: completely reported, partially reported or not reported.

The 15 items included in the OCS are as follows with definitions for each domain. Each domain can be completely reported, partially reported or not reported.

Each domain can only be completely reported, partially reported or not reported. Each domain is given a score dependent on what it is graded as.

Title

- reported completely: the title must include "randomized", "randomised", "RCT" in the title
- not reported: no report about random assignment in the title

Trial Design

- completely reported: must include the word/words "parallel", "cluster", "crossover", "factorial", "superiority", "equivalence", "noninferiority" or combinations.
- Not reported: no report of the trial design

Participants

- completely reported: eligibility criteria (health status) and location and timeframe the study was conducted are in the abstract.
- Partially reported: one eligibility criteria (health status) and location and timeframe are in the abstract.
- not reported: no report of the eligibility criteria, location and timeframe

Interventions

- completely reported: description of test and control group treatment.
- not reported: no description about treatment

Objective

- completely reported: 1 objective or primary objective clearly indicated, clearly described.
- partially reported: vague described objective or multiple ones and no primary indicated.
- not reported: no report of the objective

Outcome

- completely reported: defined primary outcome/s for the study or primary endpoint/s of the study reported
- partially reported: only 1 outcome assessed and clearly in the abstract.
- not reported: no information about primary outcome/s or endpoint/s

Randomisation

- completely reported: information in the abstract about how they randomised the participants.
- not reported: no information about the Randomisation process

Blinding

- completely reported: information about which people were blinded/masked (participants, caregivers and outcome assessors) in the abstract.
- Partially reported: use of the word/s "double-blind", "triple-blind", "single-blind", "quadruple-blind".
- not reported: no information about masking

Numbers randomised

- completely reported: must state the number of participants randomly allocated into each of the groups evident in the abstract or is easily understood.
- Partially reported: number can be added up in the abstract but is not directly reported.
- not reported: number of participants in each group is not reported and cannot be calculated

Numbers analysed

- completely reported: must state the number of participants analysed in each of the groups evident in the abstract.
- Partially reported: number can be added up in the abstract but is not directly reported.
- not reported: number of participants in each group is not reported and cannot be calculated

Outcome 2

- completely reported: reported a primary outcome, results for each group, effect size and a measure for its precision (confidence interval).
- Partially reported: reported one or more of the following items: results for each group, effect size, a measure for its precision (confidence interval).
- not reported: no data reported or only descriptive (e.g.: The survival rate at 6 months was lower in the test group. The survival rate at 12 months higher in the test group.)

Harms

- completely reported: reported which specific adverse events, side effects or complications occurred (e.g. pain, swelling, post-op bleeding, necrosis, chipping, fractures etc.) or that no such events occurred (can also be part of an outcome that is considered a measure for harms/side effects).
- Partially reported: reported about the existence of complications but did not further describe them (can also be part of an outcome that is considered a measure for harms/side effects).
- not reported: no information about side effects etc.

Conclusion

- completely reported: conclusion was stated.
- not reported: no conclusion was stated

Trial registration

- completely reported: Trial registration number was reported in the abstract.
- not reported: no information about trial registration

Funding

- completely reported: Source of funding was reported in the abstract.
- not reported: no information about funding

The OCS is calculated by taking the number of completely reported items (I_C), multiplying that by 1, taking the number of items that were partially reported (I_P), multiplying that by 0.5 and taking the items that were not reported (I_N) and multiplying that by 0, then adding together these three numbers. The maximum score is 15. Please give OCS and OCS%.

Which of the items are completely reported, partially reported and not reported and calculate the OCS% for the following abstract Detail how each domain was scored in the calculation.

'Insert abstract here'

B Let's go through each item and determine whether it is completely reported, partially reported, or not reported in the given abstract.

1. Title: Not reported (no mention of random assignment in the title).
2. Trial Design: Not reported (no information about the trial design).
3. Participants: Partially reported (mentions the number of subjects and the bilateral placement for intrasubject comparison, but does not provide information about eligibility criteria, location, and timeframe).
4. Interventions: Completely reported (provides descriptions of the synthetic bone graft and the xenograft).
5. Objective: Completely reported (states the objective of comparing histologic outcomes after sinus grafting with different materials).
6. Outcome: Not reported (no mention of primary outcomes or endpoints).
7. Randomisation: Not reported (no information about the randomization process).
8. Blinding: Not reported (no information about blinding/masking).
9. Numbers randomised: Not reported (no mention of the number of participants randomly allocated to each group).
10. Numbers analysed: Not reported (no mention of the number of participants analyzed in each group).
11. Outcome 2: Partially reported (provides results for bone-to-implant contact in both groups, but does not mention effect size or measure of precision).
12. Harms: Not reported (no information about adverse events or complications).
13. Conclusion: Completely reported (states the conclusion of the study).
14. Trial registration: Not reported (no information about trial registration).
15. Funding: Not reported (no information about funding source).

Now, let's calculate the Overall Completeness Score (OCS) and OCS% based on the scoring system provided.

I_C (completely reported items) = 4 (Items 4, 5, 12, 13)

I_P (partially reported items) = 2 (Items 3, 11)

I_N (not reported items) = 9 (Items 1, 2, 6, 7, 8, 9, 10, 14, 15)

$$\text{OCS} = (I_C * 1) + (I_P * 0.5) + (I_N * 0) = (4 * 1) + (2 * 0.5) + (9 * 0) = 4 + 1 + 0 = 5$$

$$\text{OCS\%} = (\text{OCS} / 15) * 100 = (5 / 15) * 100 = 33.33\%$$

Therefore, the Overall Completeness Score (OCS) is 5, and the OCS% is 33.33%.

Figure 1 (A) Example prompt used to generate the OCS as per CONSORT-A criteria. (B) An example of the calculated OCS and OCS% as generated by ChatGPT. CONSORT-A, Consolidated Standards of Reporting Trials for Abstracts; OCS, overall compliance score.

Table 1 Error analysis of ChatGPT CONSORT-A OCS subscores

CONSORT-A domains	Mean difference in absolute OCS	P value*	Pearson's correlation coefficient (r)
Trial design	0.065, 95% CI (0.579, 0.686)	0.054	0.49
Participants	0.228, 95% CI (0.485, 0.595)	0.001	0.26
Intervention	0.057, 95% CI (0.800, 0.881)	0.001	0.02
Objective	0.316, 95% CI (0.280, 0.384)	0.001	0.06
Outcome (methods)	0.553, 95% CI (0.077, 0.146)	0.001	0.14
Randomisation	0.633, 95% CI (0.277, 0.381)	0.001	0.11
Blinding	0.034, 95% CI (0.818, 0.895)	0.091	0.44
Number randomly assigned	0.105, 95% CI (0.530, 0.639)	0.006	0.31
Number analysed	0.028, 95% CI (0.475, 0.586)	0.434	0.04
Outcome (reporting)	0.170, 95% CI (0.453, 0.563)	0.001	0.15
Harms	0.133, 95% CI (0.602, 0.708)	0.001	0.32
Conclusion	0.764, 95% CI (0.186, 0.280)	0.001	0.06
Trial registration	0.045, 95% CI (0.918, 0.968)	0.002	0.34
Funding	0.411, 95% CI (0.533, 0.642)	0.001	0.21

*Welch's two-sample t-test.

CONSORT-A, Consolidated Standards of Reporting Trials for Abstracts; OCS, overall compliance score.

Reporting Trials for Abstracts (CONSORT-A)⁶ statement by the human authors of the study. The processes of selection and data extraction were performed independently and in duplicate by two clinician reviewers across a sample of 30 abstracts. Discrepancies were systematically addressed through discussion until a consensus of at least 80% was achieved. Subsequent data extraction was conducted solely by one reviewer. The CONSORT-A checklist scores abstract

reporting standards based on well-defined definitions for subsections such as trial design, blinding and randomisation. The human evaluators scored each item as fully reported, partially reported or not reported. ChatGPT was used to score the same set of abstracts, using a prompt to assess for each domain within the CONSORT-A checklist (figure 1). Building on the methodology established, each constituent subgroup was subsequently scored and categorised into one of the three classifications (figure 1A). An overall compliance score (OCS) was given out of 15, along with an OCS percentage (figure 1B). This was performed using the GPT3.5 model.

Bland-Altman analysis was used to evaluate the overall agreement between human and ChatGPT-generated OCS percentage. For error analysis, the mean difference of the absolute OCS subscores, Welch's two-sample t-test and Pearson's correlation coefficient were undertaken. The mean difference provides information on the magnitude and direction of the differences in OCS between ChatGPT and human evaluators, while the Pearson's correlation coefficient provides information on the strength and direction of the linear relationship between the two sets of scores. This provided complementary information on the agreement between ChatGPT and human evaluator. The Pearson's correlation coefficient was interpreted based on magnitude: r : 0–0.19 very weak, 0.2–0.39 weak, 0.40–0.59 moderate, 0.6–0.79 strong and 0.8–1 very strong correlation. Statistical analysis was done in R (V.4.1.1). $P < 0.001$ was deemed statistically significant.

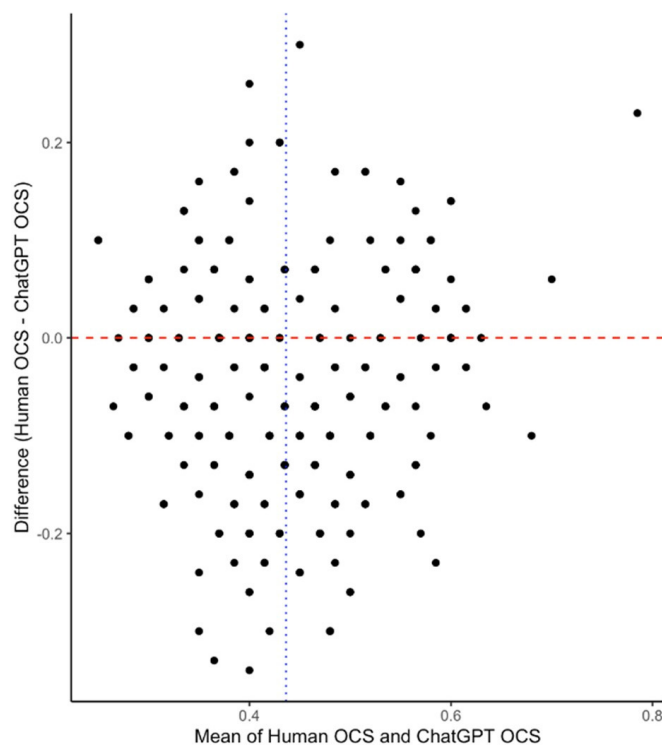


Figure 2 Bland-Altman analysis between ChatGPT human evaluation. OCS, overall compliance score.

RESULTS

Bland-Altman analysis revealed a mean difference of 4.92% (95% CI 0.62%, 0.37%) in OCS percentage (figure 2). Error analysis revealed small mean differences



between human evaluation and ChatGPT in most domains (table 1).

The mean difference in absolute OCS was highest for the 'conclusion' domain (0.764, 95% CI: 0.186, 0.280), indicating that ChatGPT differed the most from human evaluators in this domain. In contrast, the domain with the lowest mean difference in absolute OCS was 'blinding' (0.034, 95% CI: 0.818, 0.895), indicating that ChatGPT was most accurate in this domain. In terms of correlation, the study found varying levels of correlation between ChatGPT and human evaluators for different domains. For example, the domains with a strong positive correlation were 'harms' ($r=0.32$, $p<0.001$) and 'trial registration' ($r=0.34$, $p=0.002$), indicating a high level of consistency between ChatGPT and human evaluators in these domains. On the other hand, 'intervention' ($r=0.02$, $p<0.001$) and 'objective' ($r=0.06$, $p<0.001$) domains had very weak correlations, suggesting that ChatGPT's performance was less consistent with human evaluators in these domains.

DISCUSSION

The emergence of LLMs like ChatGPT offers a promising solution to streamline the assessment of reporting standards in medical literature and assist clinicians to make informed decisions. Bland-Altman analysis supports the overall findings of the study that ChatGPT has the potential to automate appraisal of medical literature. By providing a score for the quality of reporting in abstracts, ChatGPT can help clinicians and researchers quickly identify studies with more comprehensive and transparent reporting. The recent release of ChatGPT4, an advancement on the ChatGPT3 architecture, has demonstrated enhanced performance across diverse domains.^{7 8} Full access is currently limited by a paywall; however, its web integration technology creates immediate possibilities for further application. This could include searching for papers with minimum CONSORT compliance scores or the use of ChatGPT as a widget within popular medical databases, where it could automatically evaluate the quality of abstracts and provide a score to users promoting comprehensive and transparent reporting. One important barrier to using LLMs more widely in medical literature evaluation is the token limit. ChatGPT's current token limit may not allow it to process the entire research articles, limiting its use to abstracts. Nevertheless, the potential to feed ChatGPT full papers in the future and have it evaluate studies using other appraisal tools is an exciting possibility. Large, unexpected differences were seen in the conclusion and outcome (methods) subdomains. In the context of LLMs such as ChatGPT, the paucity of data in relation to training makes pinpointing a singular cause challenging. However, the quality of the prompt has been underscored as a major determinant

in response accuracy,⁹ and in the context of academic writing and interpretation, ChatGPT has been shown to not follow directions correctly.¹⁰ These may have played a pivotal role in the observed significant difference. Furthermore, some specifics of human evaluation were not elaborated upon and human assessment inaccuracies may have influenced scoring. Future research could cater to the assessment of variations between human evaluators and pave the way for a more in-depth analysis in conjunction with ChatGPT.

CONCLUSION

As the technology continues to evolve and improve, the next iteration of GPT, GPT4, may further enhance the accuracy and efficiency of the tool, allowing for even more reliable and comprehensive evaluations of research. While there are still limitations to this technology, the promise it holds for assisting in the evaluation and identification of high-quality research is a significant step towards improving patient care and outcomes.

Contributors RHRR and SRA conceptualised the study. RHRR performed the review and initial data analysis. Both RHRR and SRA were jointly responsible for subsequent in-depth data analysis. HAH, SRA, TDD and ISW contributed significantly to the editing process, refining the manuscript for clarity and consistency. All authors reviewed the final manuscript before submission.

Funding The research conducted herein was funded by Swansea University. SRA and TDD are funded by the Welsh Clinical Academic Training Fellowship (no award number). SRA received a Paton Masser grant from the British Association of Plastic, Reconstructive and Aesthetic Surgeons to support this work (no award number). ISW is the surgical specialty lead for Health and Care Research Wales and the chief investigator for the Scar Free Foundation & Health and Care Research Wales Programme of Reconstructive and Regenerative Surgery Research (no award number). The Scar Free Foundation is the only medical research charity focused on scarring with the mission to achieve scar-free healing within a generation. ISW is an associate editor for the *Annals of Plastic Surgery*, editorial board member of *BMC Medicine* and takes numerous other editorial board roles.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iD

Richard HR Roberts <http://orcid.org/0000-0002-9600-5943>

REFERENCES

- 1 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:2400.
- 2 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. 2020. Available: <http://arxiv.org/abs/2005.14165>
- 3 Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. 2020. Available: <http://arxiv.org/abs/1910.10683>
- 4 Sanmarchi F, Bucci A, Golinelli D. A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of Chatgpt using the Strobe checklist for observational studies. *Z Gesundh Wiss* [Preprint] 2023.

- 5 Menne MC, Pandis N, Faggion CM. Reporting quality of abstracts of randomized controlled trials related to implant dentistry. *J Periodontol* 2021;93:73–82.
- 6 Moher D, Hopewell S, Schulz KF, *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
- 7 He N, Yan Y, Wu Z, *et al.* Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare* 2023.
- 8 Takagi S, Watari T, Erabi A, *et al.* Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002.
- 9 Zuccon G, Koopman B. Dr Chatgpt, tell me what I want to hear: how prompt knowledge impacts health answer correctness. 2023. Available: <http://arxiv.org/abs/2302.13793>
- 10 HS Kumar A. Analysis of Chatgpt tool to assess the potential of its utility for academic writing in BIOMEDICAL domain. *BEMS Reports* 2023;9:24–30.

© 2023 Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ. <https://creativecommons.org/licenses/by/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Signal processing and machine learning algorithm to classify anaesthesia depth

Oscar Mosquera Dussan, Eduardo Tuta-Quintero, Daniel A. Botero-Rosas 

To cite: Mosquera Dussan O, Tuta-Quintero E, Botero-Rosas DA. Signal processing and machine learning algorithm to classify anaesthesia depth. *BMJ Health Care Inform* 2023;**30**:e100823. doi:10.1136/bmjhci-2023-100823

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2023-100823>).

Received 04 June 2023
Accepted 06 September 2023

ABSTRACT

Background Poor assessment of anaesthetic depth (AD) has led to overdosing or underdosing of the anaesthetic agent, which requires continuous monitoring to avoid complications. The evaluation of the central nervous system activity and autonomic nervous system could provide additional information on the monitoring of AD during surgical procedures.

Methods Observational analytical single-centre study, information on biological signals was collected during a surgical procedure under general anaesthesia for signal preprocessing, processing and postprocessing to feed a pattern classifier and determine AD status of patients. The development of the electroencephalography index was carried out through data processing and algorithm development using MATLAB V.8.1.

Results A total of 25 men and 35 women were included, with a total time of procedure average of 109.62 min. The results show a high Pearson correlation between the Complexity Brainwave Index and the indices of the entropy module. A greater dispersion is observed in the state entropy and response entropy indices, a partial overlap can also be seen in the boxes associated with deep anaesthesia and general anaesthesia in these indices. A high Pearson correlation might be explained by the coinciding values corresponding to the awake and general anaesthesia states. A high Pearson correlation might be explained by the coinciding values corresponding to the awake and general anaesthesia states.

Conclusion Biological signal filtering and a machine learning algorithm may be used to classify AD during a surgical procedure. Further studies will be needed to confirm these results and improve the decision-making of anaesthesiologists in general anaesthesia.

INTRODUCTION

Poor assessment of anaesthetic depth (AD) during general anaesthesia can result to overdosing or underdosing of the anaesthetic agent.^{1,2} In the context of anaesthetic agent overdose, extreme AD has been associated with an increased risk of mortality,³⁻⁶ intraoperative hypotension and hypoperfusion of heart and brain,⁷ perioperative nausea, vomiting and delirium.⁷⁻¹⁰ In the case of low dosage, there have been reports of intraoperative awareness, with an incidence of 0.1%–0.2%, approximately 26,000 cases per year in the USA.^{11,12}

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Poor assessment of anaesthetic depth (AD) during general anaesthesia may result in overdosing or underdosing of the anaesthetic agent. Currently, there is no integration of biological signals processed by an automatic learning algorithm that allows analysing the AD during surgical procedures and avoiding complications during surgical procedures.

WHAT THIS STUDY ADDS

⇒ A classification system has been carried out with the monitoring of brain electrical activity to assess the depth of anaesthesia. This investigation describes an AD classification process method that includes the collection of biological signals, conditioning of said signals, monitoring of the activity of the central and autonomic systems, measurement of indices and classification of patterns in AD.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This algorithm provides a reliable and well-performing tool to estimate and monitor the depth of anaesthesia in surgical procedures. The application of this innovation makes it possible to eliminate ambiguity in monitoring during the reduction of intraoperative consciousness and to reduce the risk of complications associated with deep anaesthesia.

Assessment of AD through clinical signs such as state of consciousness, limb movement, heart rate, pupil size, blood pressure, arterial blood oxygen and perspiration is used in general anaesthesia because it reflects the activity of the autonomic nervous system (ANS) and central nervous system (CNS).¹³ Evoked potentials, entropy which include state entropy (SE) and response entropy (RE), Bispectral Index and Narcotrend indices are objective measurements of the activity of the ANS.¹⁴ All these indices are based on different algorithms that analyse and record changes in electroencephalography (EEG) signals and convert them into numerical values that correspond to certain levels of unconsciousness.¹⁵⁻¹⁹ Despite quantification of anaesthetic levels by these new technologies, there are issues such as reports



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

School of Medicine, Universidad de La Sabana, Chia, Colombia

Correspondence to

Dr Daniel A. Botero-Rosas; daniel.botero@unisabana.edu.co

of ambiguity in the reduction of intraoperative awareness and burst suppression pattern misinterpretation.^{14 20–23}

Burst suppression pattern appears during deep anaesthetic levels, which may be interpreted as an error by the Bispectral Index and entropy indices^{22 24}, causing a false estimation of AD, and decreasing the safety margin between anaesthetic administration and optimal anaesthetic level.^{22 23 25–27} Another issue is that the previously mentioned indices and devices do not take into consideration ANS variables as part of the EEG indices used in DA level quantification and classification.^{14 28} Therefore, there is no definitive gold standard for the evaluation of AD levels during surgery or intensive care units.^{20 29} Regarding the evaluation of ANS activity, heart rate variability is used to determine sympathetic or parasympathetic predominance, which could provide additional information on AD monitoring during surgical procedures. In our study, a machine learning algorithm was created that uses neural networks and physiological variables to classify AD levels.

METHODS

Observational analytical study is carried out at the clinic at the Universidad de La Sabana, Chía, Colombia. Information on biological signals was collected during a surgical procedure under general anaesthesia for signal preprocessing, processing, and postprocessing to feed a pattern classifier to determine AD status of patients.

Criteria eligibility

Patients between 18 and 65 years old taken to general anaesthesia with 8-hour fasting, American Society of Anesthesiologists I and II, prior outpatient preanaesthetic assessment were included. Patients taking drugs with effects on the CNS and ANS, premedicated patients (opiates, antiemetics and sedatives such as benzodiazepines) and those who presented ANS alterations during surgical procedures, hearing and communication problems and allergy to propofol were excluded.

Data acquisition

General anaesthesia was administered with an infusion bomb using target control (B. Braun Medical, USA). Anaesthesia induction was done using 5 ng/mL of remifentanyl (Minto model) and 2.5 µg/mL of propofol (Schneider). Data acquisition was initiated 4 min before induction and finalised after having a verbal response from the patient after the surgical procedure. The EEG and ECG signals were collected using a frontal entropy sensor and the S/5TM Collect software with a sampling frequency of 300 Hz. SE and RE were collected at 0.2 Hz. The correct functioning of the non-invasive blood pressure (NIBP) sensor was also verified, and NIBP values were collected every 2.5 min. Six clinical states were defined in online supplemental file 1.

CNS signal preprocessing

The main objective is that the signal really reflects the biological phenomenon of interest, reducing artefacts

that contaminate the signal products due to electrical noise, surgical instruments and physiological artefacts such as eye movements. A technique which consisted of artefact noise filtering through a wavelet mother function was used. Those values superior to a specific threshold are removed from the signal by assigning a zero to the respective coefficient.^{30–33} Initially, 5 s of contaminated and non-contaminated EEG signal samples were selected by visual inspection of 20 records. Posteriorly, the stationary discrete wavelet transforms of six levels, with a *coiflet-3* as a mother function, was applied to each signal sample (frequency bands 0–2.33 Hz, 2.33–4.69 Hz, 4.69–9.38 Hz, 9.38–18.75 Hz, 18.75–37.5 Hz, 37.5–75 Hz, 75–150 Hz). The wavelet function (*coiflet-3*) was chosen due to its morphology and its similitude to an ocular artefact. Through observation of wavelet function (high and low frequencies) significant median differences were observed. This means that the wavelet function has the potential to treat high-frequency artefacts. Additionally, a digital filter with a cut-off frequency of 47 Hz was applied to avoid noise from the power line (50 Hz or 60 Hz), and in general terms high-frequency contamination due to surgical instrument.

An additional threshold vector for low-frequency components and a scan of low-frequency wavelet components were defined to determine significant differences between EEG epochs under general anaesthesia and epochs with contaminated EEG recordings from an awake patient (online supplemental file 2).

CNS signal processing

Complexity sample entropy (SampEn) and permutation entropy measurements were obtained from successive 5-second rectangular windows. The calculations performed for SampEn are described in online supplemental file 3. Permutation entropy provides a greater probability of prediction in general terms but fails when it must quantify the pattern associated with AD. On the other hand, SampEn provides in general terms a lower probability of prediction, but it is a good measure of complexity to predict deep anaesthesia and quantify the burst suppression pattern, prediction probability values (Pk) paired with general anaesthesia, light anaesthesia and waking state were, respectively, 0.925, 0.942 and 0.967. Permutation entropy and SampEn are combined in the proposed index as follow: permutation entropy dominates the behaviour of Complexity Brainwave Index (CBI) in the induction phase. Once the permutation entropy value crosses the median of the respective box diagram for general anaesthesia, the SampEn algorithm is activated to predict AD states. The response of the index is given according to the decision rules in online supplemental file 4.

ANS signal preprocessing

The power in the bands LF (low frequency) and HF (high frequency) was estimated using the wavelet transform, in contrast to classical methods such as Fourier analysis,

the wavelet transform does not assume stationarity of the signal analysed, and therefore fits better to evaluate transient and rapid changes in the heart rate variability series.³⁴ Wavelet Daubechies-2 was used to decompose the signal, a decomposition was performed at eight levels, the high frequency component (WC-HF) was estimated by adding the relative contribution of the coefficients of levels 4–5, and the low frequency component (WC-LF) was estimated by adding the relative contribution of levels 6–7. These values can be normalised to express proportions of a total power defined by the sum of WC-HF and WC-LF.³⁵ It is important to describe that the same wavelet filtering method was collected and applied to the ECG signal and the NIBP; later, according to the Pan-Tompkins algorithm,³⁶ R peaks were detected to form the series of relative risk intervals.

ANS signal processing

Poincare analysis and cardiac regulation: non-linear methods have been proposed to evaluate cardiac function in volunteers using pharmacological experimentation, under controlled conditions of autonomic blockade with atropine and propranolol. Two non-linear indices of autonomic function have been proposed from the Poincare descriptors: An index sensitive to vagal cardiac function called Cardiac Vagal Index (CVI), $CVI = \log_{10}(SD1 * SD2)$; an index sensitive to cardiac sympathetic function called Cardiac Sympathetic Index (CSI), $CSI = SD2 / SD1$. The change in the indices suggests a shift in regulatory activity, not the degree of activity or tone of the SNA.³⁷ The series formed by the duration of the intervals between R peaks in the ECG was analysed in windows of 60s with an overlap of 91.67%, so each time is composed of 5s of new information and the last 55s of the previous era. Initially, the classification of the patient's condition is based on the CBI indicator.

Design of pattern classifiers

The algorithms for classifying the patterns produced by the predictors of the CNS and ANS were designed with the aim of minimising the classification error in cross validation. In this way, a possible overfitting of the classifier is controlled. The classifiers were designed considering the following combinations of predictive indices {CBI, CVI}, {CBI, CSI}, {CBI, NIBP}, {CBI, CVI, CSI}, {CBI, CVI, NIBP}, {CBI, CSI, NIBP} and {CBI, CVI, CSI, NIBP}. The kth partition is used for the validation of the classification error, the classifier is adjusted or trained considering the remaining partitions of the data set. The above is done for $k=1, 2$, and finally the K classification errors are averaged. In general terms 5 or 10 partitions are recommended.³⁶

Postprocessing of CNS–ANS

The entropy parameters were postprocessed with an S-shape function (Eq. 1) to obtain a mathematical index between 0 and 100. Parameters a and b were estimated according to the values of the first awakened and third quartile deep anaesthesia of the graph on the right

in online supplemental file 4. Subsequently, a moving average filter of three entropy calculations was applied to reduce dispersion and achieve a smoother response rate that considers previous states. When a new entropy value was calculated, it was averaged with the two previous entropy calculations, or the number of entropies calculated for the first windows.

$$f(x, a, b) = 100 \left(\frac{0, x \leq a}{2 \left(\frac{x-a}{b-a} \right)^2, a \leq x \leq \frac{a+b}{2}} \frac{1 - 2 \left(\frac{x-b}{b-a} \right), \frac{a+b}{2} \leq x \leq b}{1, x \geq b} \right) \quad (\text{Eq. 4})$$

The process of classification of anaesthetic depth

This process comprised two main parts: (1) the analysis and selection of the predictors of the central nervous and autonomic systems. (2) the design of pattern classifiers. The pattern classifier was designed through the patient data set, formed by the biological signals of 60 patients (EEG, ECG, NIBP, SpO₂), and the respective anaesthesia record. Hence, the use and change of concentration of the drugs is evidenced, as well as the moment in which the patient performs some type of movement during the surgical act. The predictors' response in the following clinical events is analysed (online supplemental file 5). Clinical events define four states (categories to classify) of AD, and predictors of the CNS and ANS are described in online supplemental file 6.

Simple size and data recollection

The sample size was calculated for a correlation coefficient of 0.9, with a confidence level of 95%, accuracy of 10%, number of tests two, it is requiring a minimum of 60 subjects. Data were fully collected by the investigators and compiled using a secure server (Research Electronic Data Capture, REDCap software) and later development of the EEG Index was carried out through data processing and algorithm development using MATLAB V.8.1.

RESULTS

A total of 25 men and 35 women were included, with a total time of procedure average of 109.62 min. Regarding the EEG analysis and CBI, the results show a high Pearson correlation between the CBI and the indices of the entropy module. Nevertheless, a high Pearson correlation does not necessarily imply that the behaviour of the indices agrees. On other hand, lower correlation values were reported by the intraclass correlation coefficient between CBI and the entropy module indices. In [figure 1](#), the probability of prediction and the box diagrams corresponding to the patterns defined in the EEG. Li (light anaesthesia in recovery) and Lr (light anaesthesia on induction) were grouped in the same anaesthetic class or category, also Ak (awakened) and Rc (awakened, recovery). A higher prediction probability was provided by the CBI (Pk=0.935), SE (Pk=0.884) and RE (Pk=0.899).

A greater dispersion is observed in the SE and RE indices, a partial overlap can also be seen in the boxes associated with deep anaesthesia and general anaesthesia in these indices. A high Pearson correlation might be explained by the coinciding values corresponding to the

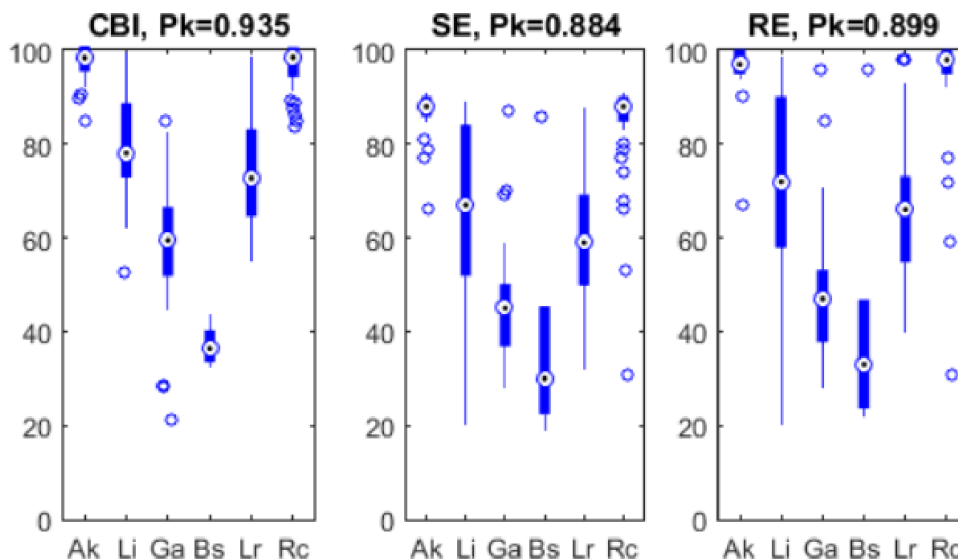


Figure 1 Box plot diagrams for EEG patterns associated with previously defined clinical states, and prediction probability values associated with CBI, SE and RE. Ak, awakened; Bs, deep anaesthesia associated with suppression burst pattern; CBI, Complexity Brainwave Index; Da, deep anaesthesia; Ga, general anaesthesia; La, light dose; Li, light anaesthesia on induction; Lr, light anaesthesia in recovery; Pmk, probability of paired prediction; Rc, awakened, recovery; RE, response entropy; SE, state entropy.

awake and general anaesthesia states. The Bland-Altman graph of [figure 2](#) shows that the differences between CBI and the entropy module indices exceed the concordance limits mainly for average values between 60 and 80 and 20 and 40, respectively. This suggests a lack of concordance in the states of light anaesthesia (estimated range: 60–80) and deep anaesthesia (estimated range: 20–40). The CBI, SE and RE associated with the defined clinical events are presented in [figure 3](#).

In the present article, we review the probability of prediction of the patient's condition was estimated for all predictors shown in [figure 4](#). In [table 1](#) La (light dose), the CBI showed a similar performance when compared

with the other indices being; SD1—light dose the best with a Pmk of 0.86, followed by CSI—light dose with Pmk of 0.85, CVI—the 0.84 Pmk and CBI 0.83.

The capacity and clinical skills of trained medical staff may be affected by external factors such as personal problems, work fatigue, among others. Besides, a physician's learning curve is not a constant independent of the previously mentioned factors, that's why it's necessary to compare the most promising machine learning methods to classify different anaesthetic levels obtaining the best outcome. In this study, the following results were obtained: In the decision tree, data set classification error and cross validation error were lowest with the data sets

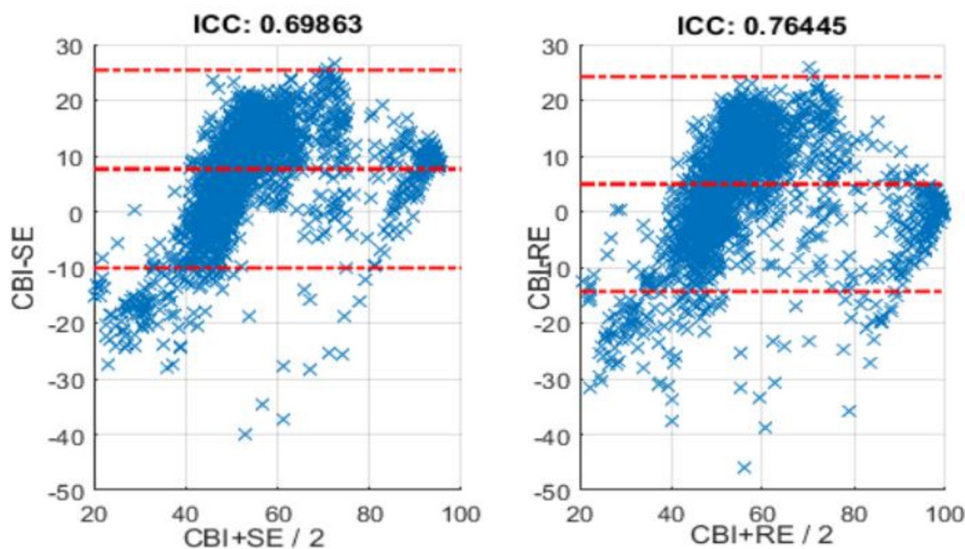


Figure 2 Bland-Altman graphs to evaluate the agreement between CBI and the SE and RE indices. CBI, Complexity Brainwave Index; ICC, intraclass correlation coefficient; RE, response entropy; SE, state entropy. *The limits of agreement are defined as the average value (red line segmented mean) \pm 2SD (red line segmented upper and lower).

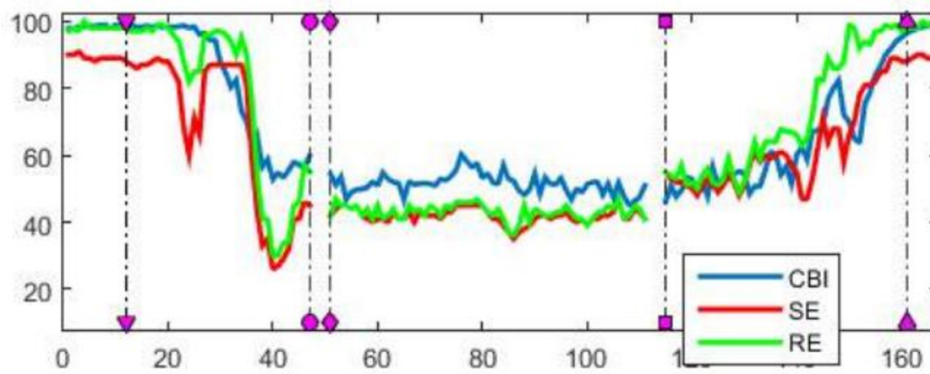


Figure 3 Values of CBI, SE and RE to different states clinical. CBI, Complexity Brainwave Index; RE, response entropy; SE, state entropy. *Triangle pointing down: induction of total intravenous anesthesia; circle: beginning of airway management; diamond: beginning of surgery; square: end of surgery; triangle pointing up: start of extubation. Figure developed by the author.

combinations of CBI–CVI–NIBP and CBI–CSI–NIBP. In the Bagging and adaptive Boosting Assembly methods, the CBI–CSI–NIBP and CBI–CVI–CSI data set groups showed the lowest classification error and X-Val errors.

In the case of the neuronal network, lowest classification error and X-Val values were in the CBI–CVI–NIBP group. On the neuro-adaptive fuzzy inference system method, the CBI–CVI data set presented the lowest errors. However,

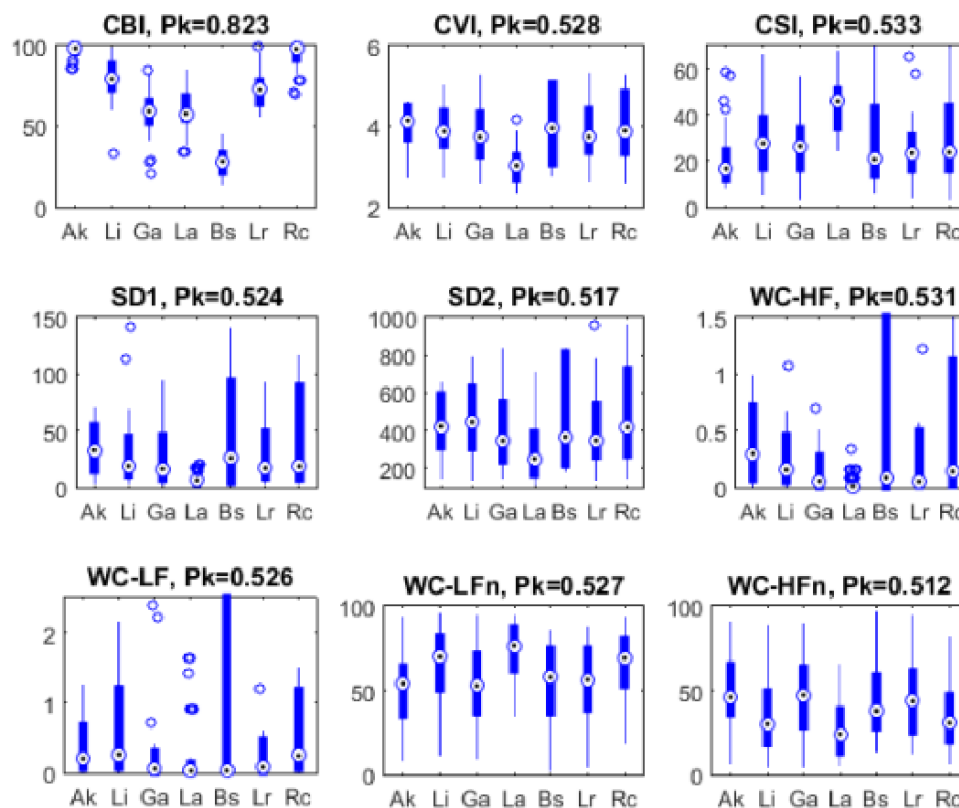


Figure 4 Box plot diagram for probability of prediction (Pk) of the patient’s condition for central nervous system and autonomic nervous system indices. Ak, awakened; Bs, deep anaesthesia associated with suppression burst pattern; CBI, Complexity Brainwave Index; CSI, Cardiac Sympathetic Index; CVI, Cardiac Vagal Index; Da, deep anaesthesia; Ga, general anaesthesia; La, light dose; Li, light anaesthesia on induction; Lr, light anaesthesia in recovery; Rc, awakened, recovery; SD1/SD2, Poincare chart descriptors; WC-HF, high frequency component; WC-LF, low frequency component; WC-HFn, high frequency power of wavelet coefficients, and respective normalisation; WC-LFn, low frequency power of wavelet coefficients, and respective normalisation. *A total of 25 light analgesia states were identified—La. There is a reduction in the performance of CBI (from 0.935 to 0.823) when considering the event light dose—La, this mainly due to overlap with the range of values associated with the event of general anaesthesia—Ga. It can be noted that SNA-related indices alone provide a poor probability of predicting the anaesthetic depth (around 0.5, which indicates that the prediction isn’t better than chance). However, the moustache diagrams seem to indicate differences in respect to other states in the methods derived from the analysis of the Poincare chart.

Table 1 Probability of paired prediction

Predictor	La	Ak	Li	Ga	Da	Lr	Rc	Pmk
CBI-La	0.97	0.89	0.53	0.98	0.69	0.98	0.83	
CVI-La	0.88	0.86	0.80	0.86	0.84	0.81	0.84	
CSI-La	0.93	0.81	0.82	0.86	0.87	0.78	0.85	
WC-HF La	0.88	0.86	0.73	0.83	0.77	0.79	0.81	
WC-HF La*	0.79	0.61	0.74	0.75	0.76	0.61	0.71	
WC-LF La	0.76	0.78	0.60	0.61	0.65	0.77	0.70	
WC-LF La*	0.79	0.61	0.74	0.75	0.76	0.61	0.71	

*These values can be normalised to express proportions of a total power defined by the sum of WC-HF and WC-LF.

Ak, awakened; CBI, Complexity Brainwave Index; CSI, Cardiac Sympathetic Index; CVI, Cardiac Vagal Index; Da, deep anaesthesia; Ga, general anaesthesia; La, light dose; Li, light anaesthesia on induction; Lr, light anaesthesia in recovery; Pmk, probability of paired prediction; Rc, awakened, recovery; WC-HF, high frequency component; WC-LF, low frequency component.

when comparing all the previously mentioned methods, the neuronal network method showed the lowest classification error and X-Val values with the CBI–CVI–NIBP (table 2).

DISCUSSION

The present study developed an algorithm that jointly considers changes in ANS and CNS pattern activity, to classify AD. Most devices used to assess anaesthetic effects on cerebral activity rely on EEG-based indices with ambiguity in reduction of intraoperative awareness.^{14 21 37} Among the most used EEG-based indices, one finds entropy and Bispectral Index.³⁸ However, there have been reports of better performance by RE Index over Bispectral Index as predictor of response to painful stimulus.³⁸ In our study, we demonstrated that an algorithm based on CBI along with other clinical variables related to ANS activity has a better performance in the classification of AD over the already known entropy indices.

Highlighting the process of innovation in medicine, we mention that this method of classification process of AD that includes the collection of biological signals, conditioning of said signals, monitoring of the activity of the

Table 2 Classifiers performance in deep anaesthesia

Predictors	Classifier	Classifier: CE	Classifier: X-Val
CBI–CVI–NIBP	Decision tree	0.086	0.118
CBI–CSI–NIBP	Decision tree	0.094	0.118
CBI–CSI–NIBP	Bagging	0.097	0.133
CBI–CVI–CSI	Boosting	0.097	0.127
CBI–CVI–NIBP	Neural network	0.094	0.103
CBI–CVI	ANFIS	0.189	0.192

ANFIS, neuro-adaptive fuzzy inference; CBI, Cerebral Brain Index; CSI, Cardiac Sympathetic Index; CVI, Cardiac Vagal Index; NIBP, non-invasive blood pressure.

central and autonomic systems, measurement of indexes and classification of patterns in AD was patented in the USA (US11504056B2), Brazil (BR112020013317A2), Colombia (CO2016002707A1) and the World Intellectual Property Organization (WO2019179544A1).^{39 40}

The main difference with the other EEG indices previously mentioned lies in the fact that this algorithm uses clinical states to classify anaesthetic states, while combining them with CNS and ANS derived predictors such as CBI, CVI, CSI and NIBP.^{41–43} The present algorithm included clinical events such as anaesthetic dose adjustment and movement during surgery as inputs in the classification of AD as a light anaesthesia state. This could explain the low global concordance between the algorithm-related CBI and the entropy indices observed in intermediate and deep anaesthesia states seen in the Bland-Altman graphs in the Results section.⁴² This means that our algorithm detects dose adjustments or movement during surgery to classify intermediate anaesthesia depth, and therefore providing more opportunities for faster detection and response in the case of intermediate anaesthesia states.⁴²

Another important aspect related to the comparison of EEG indices performance was the difference between CBI and entropy indices. A higher entropy index activity in comparison with the CBI was observed. This is most likely explained by a failure of the entropy indices in the detection of burst suppression pattern, which could be misread as the awake state.²² This could result in misinterpretation by the anaesthesiologist, which could lead in an increased anaesthetic administration. Thus, in the case of CBI, this showed a better response to burst suppression pattern. These results suggest that CBI is a better alternative; hence, reducing the error in the assessment of deep anaesthesia as the awake state and subsequent probability of dangerous anaesthetic overdose, and its derived complications.^{4 43}

In recent years, a change of paradigm has been proposed, considering the monitoring with indices based on brain electrical activity and the monitoring of standard parameters as complementary methods, and not as techniques that compete for patient care. There has been the development of classificatory system integration with other parameters correlated to ANS activity. In the present study, by comparison of cross validation errors for the different methods and a confusion matrix for neural network, different machine learning methods were implemented to estimate the best method for comparing predictors derived from CNS and ANS.⁴³

Among the different classification methods, our study found that the neuronal network with a hidden layer had the lowest cross-validation error when combining the CBI, CVI and NIBP predictors. This means our machine learning based classification algorithm had the best performance when neuronal networks were used. This clinically translates into a better prediction of AD states. However, it is important to mention the lower performance for awake and general anaesthesia states where the

highest error was seen in error matrix. Therefore, such anaesthetic states remain to be a challenge by current classificatory systems as observed in the Results section. Despite the lower prediction values, the CBI used in our algorithm still shows the highest prediction value when compared with the other predictor variables. This means our algorithm, although it presents such limitations, still performs better than the other AD classification methods. Finally, this research was based on retrospective analysis of medical records, associated with information biases; however, the research group has adequate training for the analysis and interpretation of the results. Similarly, being a single-centre study may limit the extrapolation of the results.

CONCLUSION

Biological signal filtering and a machine learning algorithm can be useful to classify AD during a surgical procedure. In our study, we show that an algorithm based on CBI together with other clinical variables related to ANS activity has a better performance in the classification of AD over the already known entropy indices.

Contributors DABR conceptualised and supervised the study and drafted, reviewed and edited the manuscript. OMD and ETQ conceptualised the study and drafted, reviewed and edited the manuscript. ETQ reviewed and edited the manuscript. ETQ reviewed and edited the manuscript. DABR drafted, reviewed and edited the manuscript, supervised the study acting as guarantor, and acquired funding.

Funding This work was supported by Universidad de La Sabana grant number MED-318-2021.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was conducted according to the Declaration of Helsinki, approved by institutional ethics committee of the Universidad de La Sabana (Acta No. 29 del 25 mayo 2012).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Daniel A. Botero-Rosas <http://orcid.org/0000-0002-7243-2238>

REFERENCES

- 1 Leslie K, Short TG. Low bispectral index values and death. *Anesth Analg* 2011;113:660–3.

- 2 Hardman JG, Aitkenhead AR. Awareness during anaesthesia. *Contin Educ Anaesth Crit Care Pain* 2005;5:183–6.
- 3 Xu Y, Shan Z, Zhao Y, et al. Association between depth of anaesthesia and postoperative outcome: a systematic review and meta-analysis. *Int J Clin Exp Med* 2018;11:3023–32.
- 4 Cui Q, Peng Y, Liu X, et al. Effect of anaesthesia de PTH on P Ostoperative clinical outcome in patients with Supratentorial tumor (DEPTH): study protocol for a randomized controlled trial. *BMJ Open* 2017;7:e016521.
- 5 Petsiti A, Tassoudis V, Vretzakis G, et al. Depth of anaesthesia as a risk factor for perioperative morbidity. *Anesthesiol Res Pract* 2015;2015:829151.
- 6 Leslie K, Short TG. Anaesthetic depth and long-term survival: an update. *Can J Anesth/J Can Anesth* 2016;63:233–40.
- 7 Cha K-M, Choi B-M, Noh G-J, et al. Novel methods for measuring depth of anaesthesia by quantifying dominant information flow in multichannel EEGs. *Comput Intell Neurosci* 2017;2017:3521261.
- 8 Leslie K, Myles PS, Chan MTV, et al. Risk factors for severe postoperative nausea and vomiting in a randomized trial of nitrous oxide-based vs nitrous oxide-free anaesthesia. *Br J Anaesth* 2008;101:498–505.
- 9 Jiang L, Nick AM, Sood AK. Fundamental principles of cancer biology: does it have relevance to the perioperative period? *Curr Anesthesiol Rep* 2015;5:250–6.
- 10 Muhlofer WG, Zak R, Kamal T, et al. Burst-suppression ratio underestimates absolute duration of electroencephalogram suppression compared with visual analysis of intraoperative electroencephalogram. *Br J Anaesth* 2017;118:755–61.
- 11 Bischoff P, Rundshagen I. Awareness under general anaesthesia. *Dtsch Arztebl Int* 2011;108:1–7.
- 12 Sebel PS, Bowdle TA, Ghoneim MM, et al. The incidence of awareness during anaesthesia: a multicenter United States study. *Anesth Analg* 2004;99:833–9.
- 13 Ha U, Lee J, Kim M, et al. An EEG-NIRS multimodal SoC for accurate anaesthesia depth monitoring. *IEEE J Solid-State Circuits* 2018;53:1830–43.
- 14 Landers R, Wen P, Pather S. Depth of anaesthesia: measuring or guessing? 2010 IEEE International Conference on Nano/Molecular Medicine and Engineering; IEEE, 2010:76–81
- 15 Musialowicz T, Lahtinen P. Current status of EEG-based depth-of-consciousness monitoring during general anaesthesia. *Curr Anesthesiol Rep* 2014;4:251–60.
- 16 Goddard N, Smith D. Unintended awareness and monitoring of depth of anaesthesia. *Contin Educ Anaesth Crit Care Pain* 2013;13:213–7.
- 17 Bithal P. Anaesthetic considerations for evoked potentials monitoring. *J Neuroanaesth Crit Care* 2014;01:002–12.
- 18 Nicolaou N, Houris S, Alexandrou P, et al. Entropy measures for discrimination of awake vs anaesthetized state in recovery from general anaesthesia. *Annu Int Conf IEEE Eng Med Biol Soc* 2011;2011:2598–601.
- 19 Kreuer S, Wilhelm W. The Narcotrend monitor. *Best Pract Res Clin Anaesthesiol* 2006;20:111–9.
- 20 Bottros MM, Palanca BJA, Mashour GA, et al. Estimation of the bispectral index by anesthesiologists. *Anesthesiology* 2011;114:1093–101.
- 21 Dahaba AA. Different conditions that could result in the bispectral index indicating an incorrect hypnotic state. *Anesth Analg* 2005;101:765–73.
- 22 Botero-Rosas DA. Monitoring the depth of anaesthesia and current technology. *JAICM* 2017;1:1–2.
- 23 Bennett C, Voss LJ, Barnard JPM, et al. Practical use of the raw electroencephalogram Waveform during general anaesthesia: the art and science. *Anesth Analg* 2009;109:539–50.
- 24 Amzica F. What does burst suppression really mean? *Epilepsy Behav* 2015;49:234–7.
- 25 Hart SM, Buchannan CR, Sleigh JW. A failure of M-entropy to correctly detect burst suppression leading to sevoflurane overdose. *Anaesth Intensive Care* 2009;37:1002–4.
- 26 Aho AJ, Lytyikäinen L-P, Yli-Hankala A, et al. Explaining entropy responses after a noxious stimulus, with or without neuromuscular blocking agents, by means of the raw electroencephalographic and electromyographic characteristics. *Br J Anaesth* 2011;106:69–76.
- 27 Bruhn J, Myles PS, Sneyd R, et al. Depth of anaesthesia monitoring: what's available, what's validated and what's next? *Br J Anaesth* 2006;97:85–94.
- 28 Shalhaf R, Behnam H, Jelveh Moghadam H. Monitoring depth of anaesthesia using combination of EEG measure and hemodynamic variables. *Cogn Neurodyn* 2015;9:41–51.
- 29 Hernandez-Meza G, Izzetoglu M, Osbakken M, et al. Near-infrared spectroscopy for the evaluation of anaesthetic depth. *Biomed Res Int* 2015;2015:939418.



- 30 Dulleck U, Ristl A, Schaffner M, *et al.* Heart rate variability, the autonomic nervous system, and neuroeconomic experiments. *J Neurosci Psychol Econ* 2011;4:117–24.
- 31 Ferreira MJ, Zanesco A. Heart rate variability as important approach for assessment autonomic modulation. *Mot Rev Educ Fisica* 2016;22:3–8.
- 32 Thayer JF, Ahs F, Fredrikson M, *et al.* A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neurosci Biobehav Rev* 2012;36:747–56.
- 33 Zikov T, Bibian S, Dumont GA, *et al.* A Wavelet based de-Noiseing technique for ocular Artifact correction of the electroencephalogram. Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society [Engineering in Medicine and Biology; IEEE, 2022:98–105
- 34 Pichot V, Buffière S, Gaspoz JM, *et al.* Wavelet transform of heart rate variability to assess autonomic nervous system activity does not predict arousal from general anesthesia. *Can J Anaesth* 2001;48:859–63.
- 35 Jeanne M, Logier R, De Jonckheere J, *et al.* Heart rate variability during total intravenous anesthesia: effects of nociception and analgesia. *Auton Neurosci* 2009;147:91–6.
- 36 Pan J, Tompkins WJ. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng* 1985;32:230–6.
- 37 Toichi M, Sugiura T, Murai T, *et al.* A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of R-R interval. *J Auton Nerv Syst* 1997;62:79–84.
- 38 Ng A: CS229 lecture notes 7.K-means; 2000. 1–30.
- 39 Botero-Rosas D, Mosquera-Dussan O. Method for classifying anesthetic depth in operations with total intravenous anesthesia. Google Patents; 2018. Available: <https://patents.google.com/patent/US11504056B2/en>
- 40 Botero-Rosas D, Mosquera-Dussan O. Proceso para clasificar profundidad anestésica en intervenciones con anestesia total intravenosa. Google Patents; 2016. Available: <https://patents.google.com/patent/CO2016002707A1/es?q=CO2016002707A1>
- 41 Avidan MS, Zhang L, Burnside BA, *et al.* Anesthesia awareness and the bispectral index. *N Engl J Med* 2008;358:1097–108.
- 42 Wheeler P, Hoffman WE, Baughman VL, *et al.* Response entropy increases during painful stimulation. *J Neurosurg Anesthesiol* 2005;17:86–90.
- 43 Schneider G, Jordan D, Schwarz G, *et al.* Monitoring depth of anesthesia utilizing a combination of electroencephalographic and standard measures. *Anesthesiology* 2014;120:819–28.

© 2023 Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Mapping loneliness through social intelligence analysis: a step towards creating global loneliness map

Hurmat Ali Shah , Mowafa Househ

To cite: Shah HA, Househ M. Mapping loneliness through social intelligence analysis: a step towards creating global loneliness map. *BMJ Health Care Inform* 2023;**30**:e100728. doi:10.1136/bmjhci-2022-100728

Received 03 January 2023
Accepted 05 September 2023

ABSTRACT

Objectives Loneliness is a prevalent global public health concern with complex dynamics requiring further exploration. This study aims to enhance understanding of loneliness dynamics through building towards a global loneliness map using social intelligence analysis.

Settings and design This paper presents a proof of concept for the global loneliness map, using data collected in October 2022. Twitter posts containing keywords such as 'lonely', 'loneliness', 'alone', 'solitude' and 'isolation' were gathered, resulting in 841 796 tweets from the USA. City-specific data were extracted from these tweets to construct a loneliness map for the country. Sentiment analysis using the valence aware dictionary for sentiment reasoning tool was employed to differentiate metaphorical expressions from meaningful correlations between loneliness and socioeconomic and emotional factors.

Measures and results The sentiment analysis encompassed the USA dataset and city-wise subsets, identifying negative sentiment tweets. Psychosocial linguistic features of these negative tweets were analysed to reveal significant connections between loneliness, socioeconomic aspects and emotional themes. Word clouds depicted topic variations between positively and negatively toned tweets. A frequency list of correlated topics within broader socioeconomic and emotional categories was generated from negative sentiment tweets. Additionally, a comprehensive table displayed top correlated topics for each city.

Conclusions Leveraging social media data provide insights into the multifaceted nature of loneliness. Given its subjectivity, loneliness experiences exhibit variability. This study serves as a proof of concept for an extensive global loneliness map, holding implications for global public health strategies and policy development. Understanding loneliness dynamics on a larger scale can facilitate targeted interventions and support.

INTRODUCTION

Loneliness is a global public health issue. Loneliness not only affects the quality of life but also leads to other mental health issues thus burdening the public health service system. Every year 162 000 Americans die from loneliness and social isolation.¹ Every forty seconds someone, around the world, commits suicide while loneliness is shown to be a direct cause of suicide.² Loneliness is shown

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Social media data are used to track mental health conditions and to gain insights into complex social and public health issues.

WHAT THIS STUDY ADDS

⇒ This paper uses social media data to understand the complex issue of loneliness which is explored in detail in social sciences but understanding it from with the help of data is lacking in literature. With the help of natural language processing tools, we analysed tweets to look for associations of socioeconomic and personal-emotional categories which are highly occurring with the mention of loneliness.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study gives insight into the dynamic and varying nature of loneliness across geographical areas. This fact can be useful for policymakers in designing specific interventions to counter loneliness by understanding loneliness in a geographical area better.

to be associated with high risk for multiple health conditions such as physical and mental health, dementia and early mortality.^{3,4} Moreover, loneliness has been shown to increase the risk of death by 26%.⁴ Loneliness is also associated with additional cost to the healthcare infrastructure. For instance, in the USA, an additional US\$6.7 billion are spent in expenses because of loneliness.⁵ Similarly, in terms of costs, loneliness costs US\$154 billion to employers in terms of absenteeism and loss in productivity.⁵

Loneliness must be understood separately from the interlinked concept of isolation. Loneliness is the subjective perception of an individual's actual and desired social connections and relationships. While social isolation on the other hand is an objective phenomenon of lack of social connections, be that with immediate family or larger community. The route to loneliness can vary from one person to the other. The relation between loneliness and social isolation with the



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

Hamad Bin Khalifa University, College of Science and Engineering, Doha, Ad-Dawahah, Qatar

Correspondence to

Mowafa Househ;
Mhouseh@hbku.edu.qa

determinant factors is complex and often bidirectional. For some people, loneliness is a prevalent state of mind. This can be the result of genetic influence or early adversity. Depression and social anxiety may lead some others to be lonely. While for some people, it may be the result of trauma and internalised stigma. These factors as well as others such as old age, economic status and negative self-image may contribute to loneliness.^{6,7} While transient loneliness can result in emotional distress, it is commonplace and can be overcome. But loneliness can become chronic and permanent because of lack of consistent and constant social connectedness, thus altering neurobiological and behavioural patterns and mechanisms.⁸

There are several intervention strategies for fighting off loneliness. These strategies are meant to mitigate the long-term mental health effects of loneliness. There have been technological interventions ranging from using social media for connectivity to videoconferencing and community-oriented interventions. The effectiveness of technology-based interventions to fight off loneliness has been studied in the literature. Authors in Döring *et al*'s study⁹ showed that communication-based technologies can reduce loneliness and isolation in older people. It was reported by Choi and Lee¹⁰ that older people using social platforms changed their behaviour through use of multifaceted technology platforms. These platforms enable social participation, cognition, nutrition and physical activity.

The recent trend in technology is towards the use of artificial intelligence (AI)-based conversational agents and chatbots. Xie and Pentina¹¹ found out through survey of patients already using a chatbot that patients form an emotional attachment with the chatbots if the patients perceive the chatbots' response to offer emotional support. The role of chatbots in interventions for mental health was studied by Boucher *et al*¹² and the potential challenges were discussed. Similarly, Abd-Alrazaq *et al*¹³ found that patients have overall feeling of satisfaction with the use of chatbot through a systematic review. However, Manis and Matis¹⁴ also pointed out the benefits of technology and chatbots particularly in terms of long-term isolation. AI-based chatbots, thus, can counter loneliness given they are complemented with other interventions.

As mentioned, digital technology interventions are shown to help in reducing the feeling of loneliness, there is a need to understand the prevalence of loneliness to devise such technology-based and community-oriented strategies. This can be understood through a loneliness map. Health informatics is applied to the area of digital health and study of loneliness through various studies. There are other studies which use social media data to gain a detailed insight into the problem of loneliness.⁹ Building on the tools of health informatics and social media analysis of mental health, digital health and loneliness a detailed global map of loneliness can act as a guideline and as the foundational grounding for intervention strategies. Loneliness is a big burden on global public health spending, global loss of accumulated number of

days of work as well as affecting the quality of life. What we need more in understanding of loneliness is from the health informatics perspective. The map, a part of which this paper will develop, will be our first towards loneliness informatics.

Through the global loneliness map, the approach is to explore the relationship between loneliness and mental health issues. This map can be used to zoom in on a country where the relationship of loneliness with negative sentiment is higher to derive further analysis. We will also provide a correlation of linguistic features representing respective personal and social categories, such as relationships, sleep habits and emotional dysregulation for different categories to show how these can vary across countries. This can be helpful in recognising and understanding the nature of association of loneliness with negative sentiments in different categories. The loneliness map will monitor the relationship of loneliness to mental health issues across the globe by analysing the data collected through ML and AI tools. The surveillance data on the relationship between loneliness and mental health issues can be used to design policy programmes to build a community of support.

This paper presents a proof of concept for such a global loneliness map. Developing the loneliness map which is exhaustive and backed by rigorous evidence is a time and resource intensive project. This paper presents the first step towards it. The remaining parts of the map, that is, using multiple data sources and analysing different regions and countries exhaustively will be carried out stepwise. Rather than using multiple sources of data we first focus on Twitter because the data it provides is diverse as well as from a limited dataset multiple insights can be gained as the users have to express themselves in limited characters. Moreover, we start with the USA. We collected data mentioning keywords associated with loneliness and found out that the data returned by the Twitter algorithm has more tweets from the USA. We collected global data on loneliness as we wanted a snapshot into loneliness rather than exhaustive analysis of one country. We retrieved the US cities which have more than 10 000 tweets each related to loneliness.

To develop the first part of loneliness map, we used sentiment analysis of Twitter data through natural language processing tool. This is based on psycholinguistic model of understanding mental health issues. The collected tweets are stored in a database and then sentiment analysis using valence aware dictionary for sentiment reasoning (VADER)¹⁵ tool from the natural language toolkit (NLTK) is carried out. VADER is lexicon and rule-based model for sentiment analysis. The lexicon-based approach means that the algorithm is constructed using a dictionary which contains a detailed list of sentiment features. In addition, VADER also complements the lexicon-based dictionary with grammatical rules which are heuristic in nature. These rules complement the lexicon-based sentiment analysis to determine polarity of the sentiment. The result of the sentiment analysis

tool gives us an indication of loneliness in the particular dataset.

LITERATURE REVIEW

Understanding loneliness theoretically and its relation to mental health has been the subject of several studies such as.^{10 16 17} From the health informatics side, there also have been studies which deal with the application of technology-based intervention to cope with loneliness such as.⁹ Loneliness is shown by these studies to be associated with increased risk of mental health issues. Interventions for loneliness which are based either on technology or through building community were shown to be effective in reducing the negative effects of loneliness.

Technology is used to fill in the gap created by lack of access to a healthcare professional or service. Byrne *et al*¹⁸ carried out a scoping review of reviews to study the effectiveness of communication technologies to reduce the feeling of loneliness in older people. The study concluded that communication-based technologies do in-fact reduce feeling of loneliness in older people. Similarly, Hards *et al*¹⁹ studied through a systematic review and meta-analysis of digital technologies-based intervention to reduce loneliness in older adults. The study analysed 6 articles finally with 646 participants reported in combined. The study showed no statistical difference between the effectiveness of digital intervention, but it self-reported the lack of enough studies and small sample size of participants to be the cause for lack of validating effectiveness.

However, Döring *et al*⁹ establish the relationship between communication-based technologies and reduction in the feeling of loneliness. Through a cross-sectional study of 4315 older adults, aged above 50, it reported that rural older adults who used technology less-frequently felt loneliness more than urban older adults. Choi and Lee¹⁰ carried out a study of effectiveness of social networking sites usage in older people for reducing loneliness. The study found some evidence that the use of social networking sites was associated with reduction in feeling of loneliness and reduction in feeling of depression. But the studies lacked on the experimental side.

The brief literature review provided above provides the scientific foundation for effectiveness of technology-based intervention in loneliness. However, there is a gap in global understanding and prevalence of loneliness. Surkalim *et al*²⁰ carried out a study of prevalence of loneliness in 113 countries to identify data availability, gaps

and patterns for population level existence of loneliness. However, the study did not design a tool, nor an intervention based on the meta-analysis carried out.

Twitter has been used for studying other phenomena and public health concerns such as.^{21 22} Data were collected in Guntuku *et al*²³ from twitter to study loneliness. Twitter is also used for other mental health related topics such as a detailed study of tweets related to insomnia and its correlation with mental health was carried out by Maghsoudi.²⁴ Similarly, Alhuzali *et al*²⁵ carried analysis of emotions in the UK and geo-located the emotions across different cities to find the sentiment during COVID-19 pandemic. For mental health problems²⁶ carried analysis of twitter data to detect the magnitude of depression. Given the literature overview, the following are the scope and contribution of this paper:

1. This study provides proof of concept for a loneliness map where the dynamics of loneliness can be understood through publicly available social media and other online data.
2. How the topics and themes associated with loneliness over Twitter relate to larger socioeconomic and personal-emotional categories?
3. Is there a difference in aggregate expression of loneliness, thus pointing to the dynamic nature of loneliness, even across the same country or does the expression change across the country relevant to different geographical or socioeconomic conditions?

DATA PROCESSING AND SENTIMENT ANALYSIS

In this section, we will present how the data are collected, discuss the data sources as well as sentiment analysis carried out on the data.

Methodology

The study does not use identity of persons involved generating the data but gives an aggregate and an overall picture based on opinions expressed publicly. We use social intelligence analysis (SIA) to find the correlation of loneliness with mental health problems and other correlated topics. The SIA is a broad theme which incorporates multiple social media sources such as Facebook, Reddit and Quora, etc. SIA is important to gain insight into user's data and in our case understand the dynamics of loneliness. While SIA can be used for a variety of purposes such as mining content to create stories or to find out trends, we have used SIA for sentiment analysis of collected data on loneliness. As mentioned in the introduction, this is the proof of concept or first step towards a global loneliness map. Therefore, we have only used Twitter for collecting data and used a sentiment analysis tool for analysing the sentiment of data retrieved from USA, which mentions keywords associated with loneliness.

We used respective analysis of publicly available data of users posting about loneliness. Twitter is a social media platform which is used for connectivity and opinion sharing and allows users to post via short messages

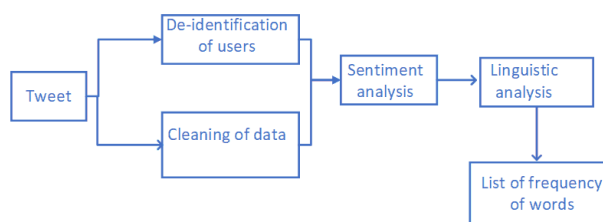


Figure 1 Pipeline for processing Twitter data.

consisting of 280 characters. Twitter gives access to the users' data through its publicly available Twitter API for developers. The data we gathered was based on topic modelling through open-vocabulary topics. The relevant tweets about loneliness were gathered and stored in a database. Topics, which are combinations of clusters of co-occurring words were created. These topics are then analysed further through a dictionary-based approach. Our approach also relies on dictionary-based psycholinguistic features to create a loneliness map as is used by Pennebaker *et al.*²⁷

For topic modelling, we used the words 'lonely', 'loneliness', 'alone', 'isolated' and 'isolation' to give a list of tweets containing these keywords. In theoretical literature, the words 'loneliness' and 'lonely' are used to describe the feeling under consideration in this paper. Authors in Guntuku *et al.*'s study²³ collected Twitter data based on keywords 'lonely' and 'alone'. We went further and included the synonyms and related words with loneliness for collecting our Twitter data.

We did not want to exhaustively search for one specific country because we wanted the data collected to be proof of concept. We can focus exhaustively on the cities or countries and collect more data about them based on the data collected in this step. The data collected were analysed through a sentiment analysis approach to find out the topics most correlated with loneliness in different cities in the USA. The next subsection explains why sentiment analysis on the collected data is needed.

Sentiment analysis

We collect a particular number of tweets with the keywords for loneliness. If we were reporting all the tweets that contained feelings of loneliness, we would not have required a further step. In that case, the problem becomes determining the association or correlation between themes (which may represent loneliness) with the keywords depicting loneliness. For instance, we had to find what is the relationship between 'hurt', 'sick', 'tired', 'sleep', etc with the expression of loneliness. This task is usually carried out by association of lexicon categories with tweets including the words 'lonely' or 'alone'.

The problem we are formulating in this paper is on a larger scale. Thus, the limited scale of representative tweets has to be interpreted in a novel way to give us any meaningful insight into loneliness. All the tweets in each dataset contain keywords representing loneliness. These data can be analysed in one way to give association between loneliness with other categories across the globe for different selected countries. This trend in its own is important to give a global picture of determinants of loneliness and to give a tool to policy-makers to address loneliness in their specific country. But the mention of 'lonely' or 'alone' can also be in a non-negative way. This fact gives us an opportunity to look at the relationship between mentioning keywords representing loneliness and negative emotions which may ultimately be linked to psycholinguistic feature of mental well-being.

For establishing the correlation between loneliness and negative sentiment we used VADER based on Python's NLTK. VADER is suited for microblog content, such as that of Twitter. VADER combines lexicon, that is, dictionary-based analysis, and rule-based approach to characterise the sentiment. Other lexicon-based sentiment analysers such as linguistic inquiry and word count (LIWC)²⁷ are only polarity based. VADER on the other hand also gives valence of the sentiment on the range from 1 to 9. Because of the sentiment score we can also know through VADER the extent to which the sentiment is negative or positive.

This valence is based on generalisable rules that represent grammatical and syntactical conventions that humans use in contexts meant for emphasising a sentiment intensity.

For our purposes, another important feature of VADER is the inclusion of sentiment bearing lexical non-verbal items such as emoticons and verbal items such as slang, acronyms, initialisms which are prevalent in social media context. The combination of valence polarity though both lexicon and rule-based approach are valuable for fine-grained sentiment analysis. VADER overcomes the shortcomings of lexicon-based analysers such as LIWC through a machine learning approach. The shortcomings of lexicon-based approach come in coverage, general sentiment intensity and acquiring a new set of human lexical features.

In this paper, through the global loneliness map, the approach is to correlate the categories of loneliness with possible negative mental health outcomes. This map can be used to zoom in on a country where the relationship of loneliness with negative sentiment is higher to derive further analysis. We also provide a correlation of linguistic features representing respective personal and social categories, such as relationships, sleep habits and emotional dysregulation for different categories to show how these can vary across countries. This can be helpful in recognising and understanding the nature of association of loneliness with negative sentiments in different categories. Subsequently, this can guide intervention strategies in those specific areas.

RESULTS

Data about the keywords associated with loneliness were collected during October 2022 through the developer API of Twitter. The purpose of this paper is not to find the number of people with loneliness in a particular area or country. That kind of study would require collecting billions of tweets. Rather the purpose in this study is to find the correlations of loneliness with socioeconomic, political and personal-psychological categories. For this purpose, we do not need to go deeper into a user's timeline and monitor their activity. We are more interested in the aggregate behaviour of users in relation to the expression of loneliness. We deidentified the tweets before analysing them, that is, we remove the users' names and

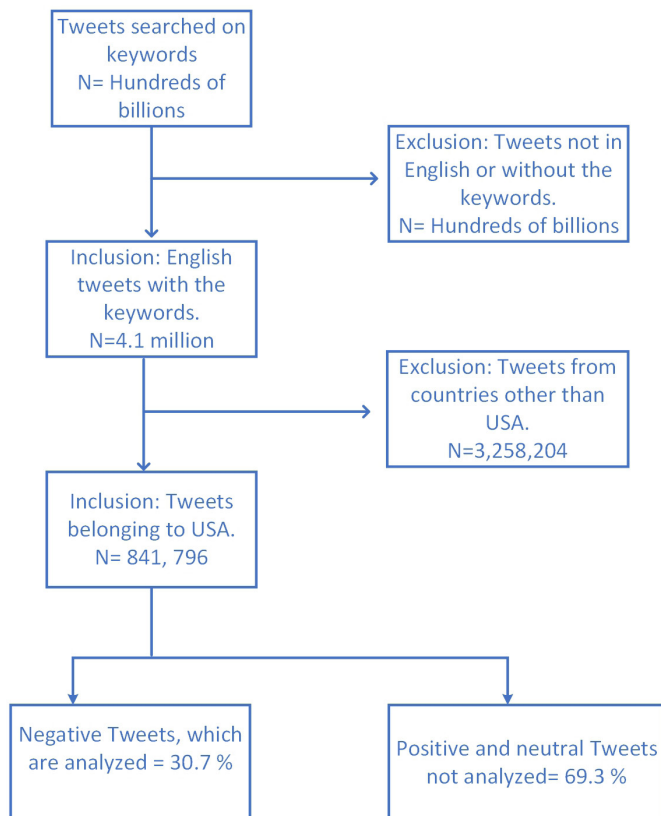


Figure 2 Strengthening the Reporting of Observational Studies in Epidemiology diagram for the Twitter data.

IDs. This is part of the data cleaning process. The data are publicly available, but we will not disclose the collected data without anonymising it.

Globally, 4.1 million tweets were collected. Out of these 841 796 were from the USA. Five cities had tweets higher than 10000 which we analysed. We also analysed one city with tweets less than 10000 but higher than 5000 to see whether the result conforms to the other cities with number of tweets more than 10000. Orlando was the city, and the number of tweets was 5535.

Figure 1 presents our pipeline of analysis of data collected from Twitter. Twitter gives access to the users’ data through its publicly available Twitter API for developers. The data we gathered was based on topic modelling through open-vocabulary topics. The relevant tweets about loneliness were gathered and stored in a database. Topics, which are combinations of clusters of co-occurring words, were created. These topics are then analysed further through a dictionary-based approach.

Tweets were collected containing the keywords mentioned in the last subsection. Tweets were extracted from these two countries to make a subdataset belonging to the USA. This was meant to reflect the majority composition of the dataset. Sentiment analysis was carried out after cleaning the data such as removing redundant characters, numbers, special characters, users’ profile ID and information such as ‘retweet’. Sentiment analysis is important to differentiate between the phrases and topics carrying meaningful information on loneliness

Table 1 Sentiment analysis of tweets containing the keywords/topics of loneliness

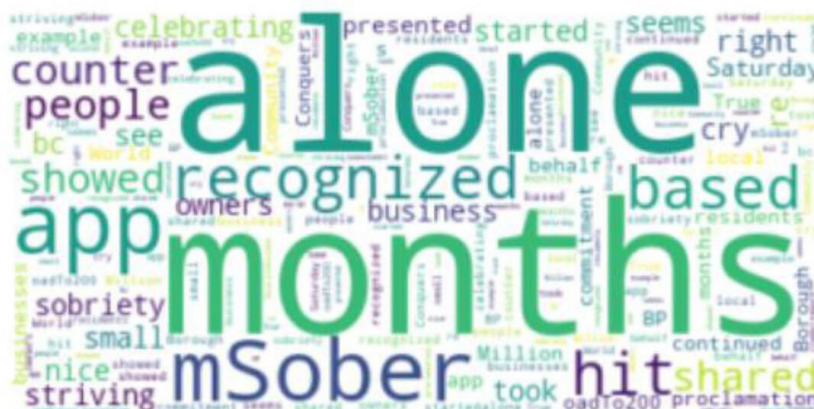
City	Sentiment analysis of tweets
Houston	Positive: 9.6% Negative: 21.2% Neutral: 69.2%
Nashville	Positive: 48.5% Negative: 51.5% Neutral: 0.0%
Orlando	Positive: 52.6% Negative: 47.4% Neutral: 0.0%
Queens	Positive: 19.4% Negative: 80.6% Neutral: 0.0%
San Francisco	Positive: 46.7% Negative: 47.3% Neutral: 6.0%
Washington	Positive: 40.8% Negative: 56.5% Neutral: 2.7%
USA	Total tweets: 841 796 Positive: 139 210, 16.5% Negative: 258 401, 30.7% Neutral: 44 185, 52.8%

and metaphorical and non-sequitur uses of the terms and topics associated with loneliness. Figure 2 gives the process of collecting data from Twitter and the process of analysis of the tweets.

Table 1 gives sentiment analysis for different cities as explained above and for the overall dataset which contains data about the USA. Table 1 also points towards an interesting outlier in the dataset, that is, Houston accounts for almost all the neutral tweets. Some of the cities have a more balanced amount of negative and other tweets (ie, positive and neutral) while two clear outliers can be pointed out in the dataset. For Houston, only 21.2% tweets are negative while for Queens 80.6% tweets are negative. The data were collected for 2 weeks, and it is not wide and deep enough to know with certainty the causes of these outliers. As mentioned, this study is a proof of concept for a wider loneliness map on the basis of SIA, that is, through analysing various social media and web based data through the tools of machine learning and AI. However, the neutral tweets along with the positive tweets do not add to the analysis of loneliness as carried out in this paper. With this dataset, the reason for these outliers cannot be ascertained without looking further into long-term data for each city. In further studies, the long term data will be collected to have balanced dataset for each city and find out the reasons for proportion of each category of tweets.

The aim of the loneliness map and this paper is to find the correlation between loneliness and mental health issues and other topics which can vary from personal

A



B



Figure 3 Words more likely to be posted by Twitter users (A) when the sentiment of the tweet is positive, (B) when sentiment of the tweet is negative.

expression to socioeconomic factors. Before going into detailed analysis of the tweets on loneliness, it was important to find out the tweets which are metaphorical or non-sequitur. The neutrality can also represent the mention of loneliness in descriptive terms. The data here show that the sample size is consistent in producing reliable results as Orlando with the smallest sample size has similar results as other cities.

Figure 3 presents the word clouds of the sentiment of the tweets. This figure illustrates the most highly associated words with the groups of users tweeting with keywords associated with loneliness. It is important to plot the word cloud of both positive tweets and negative tweets to differentiate between metaphorical use and the meaningful use as intended by the study design of this paper. From the figure it can be seen that the words associated with positive sentiment of mention of loneliness are positive words such as commitment, sobriety, sober and months (number of months). The word cloud was generated after redundant words were removed such as the 'RT' (retweeted) and mention of the user's ID.

Table 2 presents the highly correlated topics with negative mention of loneliness. The tweets with negative sentiment were first tokenised and stemmed to get a concise

list of words and topics associated with loneliness. The list was then analysed and meaningful words representing topics of interest such as emotional, social and health, etc identifiers were found out. Words such as 'oh', 'yeah' and 'ur' were ignored in composing the list. From table 2, it can be seen for the overall US dataset intimate relationships followed by interpersonal relationships are the highest correlated topics, thus, issues associated with loneliness. 'COVID-19' is the single highest occurring word in the dataset. The search keywords contained 'isolated' and 'isolation' and given the social and physical distancing required by COVID-19 prevention guidelines the highest occurrence of COVID-19 in association with negative sentiment of loneliness is expected. This tells us that the isolation because of COVID-19 has negative effects on people's sentiments, thus their overall mental health. We also found the association of drug and addiction words with loneliness. The same was also found in figure 3B where the word 'sober' which is associated with recovery from addiction was used although in a positive sense. The combination of both figure 3B and table 2 shows the association of drug/alcohol addiction with loneliness; thus, it can be further investigated with keywords associated with both loneliness and addiction.

Table 2 Highly correlated topics with mentions of loneliness

Highly correlated topics with negative mentions of loneliness. Topics are divided into a broader theme area		
Thematic area	Topic	No. of mentions
Intimate Relationships	Cheat	13 395
	Man	8250
	Family	7196
	Woman	5941
	Relationship	4926
	Marriage	3736
Interpersonal Relationships	Want	13 655
	Need	12 868
	Feel	11 018
	Hurt	1680
	Forgot	1104
Health	Covid	11 313
	Die	8972
	Life	5741
	Patient	1210
Socio-economic factors	Colorism (Black/White)	27 740
	Money	5179
	Poor	3391
	Racism	1230
	Slavery	1197
Emotional expression/insecurities	Sad	8942
	Hate	6581
	Fat	3883
	Anger	3525
	Sexual	3360
	Grieve	2115
Drug/Alcohol	Rehabilitation	8580
	Smoke	3554
	Drunk	959
Insomnia	Night	4245
	Awake	1269
	Sleep	1204
	Bed	627

In [table 3](#), the city-wise topic association of themes with loneliness was found out. It was based on analysis of tweets with negative sentiment. It was found out that the sample, however, limited, contained variation as per themes and topics associated with the negative consequences of loneliness. While these topics and their association with loneliness are not definitive, that is, it may change with availability of more data per city, it provides proof of concept for the idea of mapping loneliness, nonetheless. Some of the correlations are intuitive and

Table 3 Top correlated topics with negative mention of loneliness across cities analysed

City	Top three correlated topics
Houston	1. Cheat 2. Family 3. Relationship
Nashville	1. Black, white (mention of racism/colorism) 2. Missing someone 3. Addiction
Orlando	1. Racism 2. Emotional expression (hate) 3. Sexual identity (straight)
Queens	1. Self-focused (mention of 'self') 2. Covid 3. Emotional expression (forgot)
San Francisco	1. Emotional expression/dysregulation (Sh*t, F*ck, Broke) 2. Addiction 3. Black, white (mention of racism/colorism)
Washington	1. Rehabilitation 2. Emotional expression/dysregulation (Sh*t, F*ck, Eat) 3. Health (Mask, Ebola)

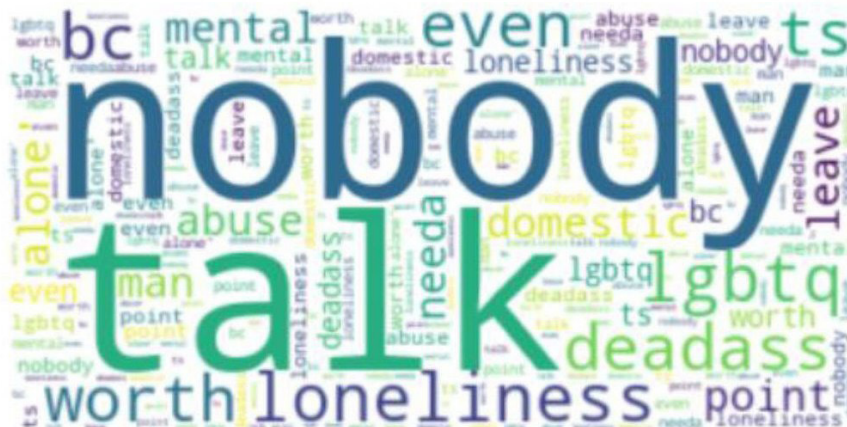
self-expressive, for example, Queens being a big city with the peculiar nature of big city one would expect more self-oriented or self-focused expressions. The data analysed here provide a peek for data collected over a limited period, but it proves that the expression and dynamics of loneliness can change with geography which in turn can be dependent on particular urban infrastructure, health-care system, socioeconomic issues and culture of the region. Similarly, [figure 4](#) shows a few selected examples of city-wise association of topics with loneliness. As can be seen each word cloud is different with some meaningful words contained in each. For example, in Houston the word 'lgbtq' can be seen, while for Orlando words such as 'love' can be spotted out. This again drives home the point of variance in experience and expression of loneliness. It must be noted that the word cloud is based on the full words and phrases while the list in [table 2](#) is based on stemmed words.

DISCUSSION AND LIMITATIONS

The methodology developed in this paper shows the association of loneliness with language which is associated with mental health issues such as anger and depression. The tweets analysed prove that psychosocial linguistic features can be found in self-expression of loneliness which can identify dynamics of loneliness.^{28–30} Further, we present the topics and themes associated with loneliness can vary along both the thematic area and the geographic region. Tweets containing keywords associated with loneliness



A Houston



B Orlando



C Nashville

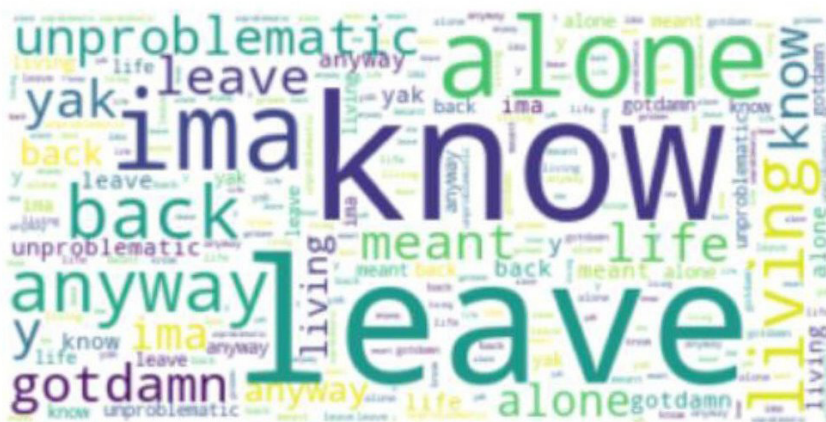


Figure 4 Selected city-wise examples of world clouds of words/topics associated with negative sentiment of loneliness, (A) Houston, (B) Orlando, (C) Nashville.

also represent a self-focused discourse which affirms previous literature on loneliness.^{31 32} Tables 2 and 3 also point towards other results which have been established in literature on loneliness. These include conformity with literature on association of loneliness with substance abuse, emotional dysregulation and trouble with relationships.³³ A loneliness map developed by SIA through

machine learning and social media data analysis can thus be a powerful tool for policy-makers.

As mentioned in the Literature review section, there are very rare studies carried out on studying loneliness through Twitter, therefore, this paper is a novel idea in studying and understanding loneliness. However, Twitter and social media have been used to study other mental

health and public health concerns. Loneliness was studied in detail for Pennsylvania by Guntuku *et al.*²³ The study provides insight into how loneliness is felt in the particular region. Our study goes beyond this study and carries out a comparative analysis of loneliness across six cities of the USA. Moreover, this study will be used as a proof of concept for a detailed map of loneliness on a global scale. Twitter data were used by Melton *et al.*³⁴ to study the response to vaccination against COVID-19. They developed their own sentiment analysis model, but they did not provide detailed analysis of the users' tweets. The categorisation of socioeconomic, political and personal-psychological topics was missing from the study which this study provides. Similarly, the dynamics of insomnia and its correlation with different external and internal factors was not carried out by²⁴ as compared with this study which goes in depth to give the dynamics of the topic of study, that is, loneliness.

There are some limitations of this study. The first limitation is that the dataset size is small as compared with the actual data being generated by both countries on the keywords of loneliness. Tweets can run into millions even for a city on the keyword of loneliness. But the purpose of this study is not to carry out a rigorous analysis but to give a proof of concept for a loneliness map. The other limitation of this study on another front is the automatic classification of Tweets into negative and positive through sentiment analysis. While this has been the basis of the paper to carry out automated analysis, the result of this automated sentiment analysis needs to be validated through looking at a certain number of Tweets which have been identified negative. Through this way, we will be able to know the confidence of analysis and quantify the error.

CONCLUSION

This paper develops the proof of concept for loneliness map project. In this proof of concept, we analysed different cities in the USA through data collected from Twitter to see the correlation of loneliness with negative sentiment and other correlated topics. The loneliness map will be incrementally developed by considering multiple data sources and different regions and countries of the world. The loneliness map project will integrate multiple data sources (such as social media content, surveys and news) to analyse loneliness through ML and AI to create a map of loneliness across the globe to understand the impact of loneliness on mental health. Loneliness map is not only meant to see the prevalence of loneliness in different countries, regions and cities around the world, but it will also be instrumental in understanding the impact of different sociocultural, political, economic and geographical dynamics on loneliness and mental health. The loneliness map can guide intervention for policy-makers in healthcare such as the health map in.³⁵ The interventions can also be guided by data provided by the loneliness map. The division between urban and rural,

economic zones and classes and their relationship with loneliness can be observed from a loneliness map. The map can also trace historical data of loneliness in particular regions and find the co-relationship of increasing loneliness with mental health. From the digital mental health perspective, the question that whether loneliness is the result of mental health problems or cause, can be answered through the data provided by the loneliness map.

In this paper, sentiment analysis of tweets containing keywords associated with loneliness was carried out for the US cities. The results showed variance in the sentiment associated with loneliness in different cities as well as the top correlated topics with the mention of loneliness. This can be important for policy-makers to understand the particular nature of loneliness in these cities. These results are only indicative and will need further exhaustive study. To point out for the sake of clarity, the number of tweets containing the keywords associated with loneliness can run up to millions for a particular city during a year. But the objective of this paper is not to study exhaustively each city but to determine from the data collected the sentiment associated with loneliness in order to prove that the dynamics of loneliness are not the same even in the same country. This provides a peep into the varying nature of loneliness, thus driving the point home that loneliness can be varied and would need different strategies to counter the negative feelings associated with loneliness.

In future, this work can be extended in many directions. We plan to extend the analysis with the same cities by collecting focused data and analysing it in more detail to find the socioeconomic and personal-emotional dynamics of loneliness for the city. We also will collect data about loneliness from different countries in different languages, translate the data and analyse through sentiment analysis. In further work in data collection, we will use other social media platforms such as Facebook, Reddit and Quora to collect data and topics on loneliness. The data on these platforms are detailed which can give more intimate analysis of a person's experience of loneliness. While Twitter's data are diverse, the data of these other social media platforms would be detailed and intimate. The data from these different social media platforms can then be combined to have a more accurate understanding of loneliness, thus also improving the quality of the loneliness map.

Contributors HAS designed the study, carried out the analysis and wrote the paper. HAS is also guarantor of the study. MH designed and supervised the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval No ethical approval was sought for this study as this study analyses publicly available online content.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Hurmat Ali Shah <http://orcid.org/0000-0002-5691-4819>

REFERENCES

- 1 An epidemic of loneliness and despair: how wisdom can help. n.d. Available: <https://www.bbrfoundation.org/blog/epidemic-loneliness-and-despair-how-wisdom-can-help>
- 2 Shaw RJ, Cullen B, Graham N, *et al*. Living alone, loneliness and lack of emotional support as predictors of suicide and self-harm: A nine-year follow up of the UK Biobank cohort. *J Affect Disord* 2021;279:316–23.
- 3 Shah SGS, Nogueras D, van Woerden HC, *et al*. Evaluation of the effectiveness of Digital technology interventions to reduce loneliness in older adults: systematic review and meta-analysis. *J Med Internet Res* 2021;23:e24712.
- 4 Peytrignet S, Garforth-Bles S, Keohane K. Loneliness Monetisation report: analysis for the Department for Digital, culture, Media & Sport 2020. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/94482/loneliness-monetisation-report.pdf
- 5 How loneliness is damaging our health. n.d. Available: <https://www.nytimes.com/2022/04/20/nyregion/loneliness-epidemic.html>
- 6 Wister A, Fyffe I, O'Dea E. Technological interventions for loneliness and social isolation among older adults: a Scoping review protocol. *Syst Rev* 2021;10:217.
- 7 Pitman A, Mann F, Johnson S. Advancing our understanding of loneliness and mental health problems in young people. *Lancet Psychiatry* 2018;5:955–6.
- 8 Hickin N, Käll A, Shafran R, *et al*. The effectiveness of psychological interventions for loneliness: A systematic review and meta-analysis. *Clin Psychol Rev* 2021;88:S0272-7358(21)00109-4.
- 9 Döring N, Conde M, Brandenburg K, *et al*. Can communication Technologies reduce loneliness and social isolation in older people? A Scoping review of reviews. *Int J Environ Res Public Health* 2022;19:11310.
- 10 Choi HK, Lee SH. Trends and effectiveness of ICT interventions for the elderly to reduce loneliness: A systematic review. *Healthcare* 2021;9:293.
- 11 Xie T, Pentina I. Attachment theory as a framework to understand relationships with social chatbots: a case study of replika. Hawaii International Conference on System Sciences; 2022
- 12 Boucher EM, Harake NR, Ward HE, *et al*. Artificially intelligent Chatbots in Digital mental health interventions: a review. *Expert Rev Med Devices* 2021;18:37–49.
- 13 Abd-Alrazaq AA, Alajlani M, Ali N, *et al*. Perceptions and opinions of patients about mental health Chatbots: Scoping review. *J Med Internet Res* 2021;23:e17828.
- 14 Manis KT, Matis J. AI companionship or loneliness: how AI-based Chatbots impact consumer's (Digital) well-being: an abstract. In Academy of Marketing Science Annual Conference-World Marketing Congress; Cham: Springer International Publishing, 2021:365–6
- 15 Hutto C, Gilbert EV. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *ICWSM* 2014;8:216–25.
- 16 Gierveld J de J. A review of loneliness: concept and definitions, determinants and consequences. *Rev Clin Gerontol* 1998;8:73–80.
- 17 Wiwatkunupakarn N, Pateekhum C, Aramrat C, *et al*. Social networking site usage: A systematic review of its relationship with social isolation, loneliness, and depression among older adults. *Aging & Mental Health* 2022;26:1318–26.
- 18 Byrne KA, Anaraky RG, Dye C, *et al*. Examining rural and racial disparities in the relationship between loneliness and social technology use among older adults. *Front Public Health* 2021;9:723925.
- 19 Hards E, Loades ME, Higson-Sweeney N, *et al*. Loneliness and mental health in children and adolescents with Pre-Existing mental health problems: A rapid systematic review. *British J Clin Psychol* 2022;61:313–34. 10.1111/bjc.12331 Available: <https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bjc.12331>
- 20 Surkalim DL, Luo M, Eres R, *et al*. The prevalence of loneliness across 113 countries: systematic review and meta-analysis. *BMJ* 2022;376:e067068.
- 21 Ali F, Ali A, Imran M, *et al*. Traffic accident detection and condition analysis based on social networking data. *Accident Analysis & Prevention* 2021;151:105973.
- 22 Zeberga K, Attique M, Shah B, *et al*. A novel text mining approach for mental health prediction using bi-LSTM and BERT model. *Comput Intell Neurosci* 2022;2022:7893775.
- 23 Guntuku SC, Schneider R, Pelullo A, *et al*. Studying expressions of loneliness in individuals using Twitter: an observational study. *BMJ Open* 2019;9:e030355.
- 24 Maghsoudi A, Nowakowski S, Agrawal R, *et al*. Sentiment analysis of insomnia-related Tweets via a combination of transformers using Dempster-Shafer theory: pre- and peri-COVID-19 pandemic retrospective study. *J Med Internet Res* 2022;24:e41517.
- 25 Alhuzali H, Zhang T, Ananiadou S. Emotions and topics expressed on Twitter during the COVID-19 pandemic in the United Kingdom: comparative Geolocation and text mining analysis. *J Med Internet Res* 2022;24:e40323.
- 26 Stephen JJ, P. P. Detecting the magnitude of depression in Twitter users using sentiment analysis. *IJECE* 2019;9:3247. 10.11591/ijece.v9i4.pp3247-3255 Available: <http://ijece.iaescore.com/index.php/IJECE/issue/view/487>
- 27 Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 2001;71.
- 28 Stravynski A, Boyer R. Loneliness in relation to suicide Ideation and Parasuicide: A Population-Wide study. *Suicide Life Threat Behav* 2001;31:32–40.
- 29 Blai B. Health consequences of loneliness: A review of the literature. *J Am Coll Health* 1989;37:162–7.
- 30 Hawkey LC, Cacioppo JT. Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Ann Behav Med* 2010;40:218–27.
- 31 Edwards T, Holtzman NS. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality* 2017;68:63–8.
- 32 Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its relationship to feelings of loneliness. *Computers in Human Behavior* 2014;36:460–8.
- 33 Booth R. Toward an understanding of loneliness. *Social Work (Stellenbosch)* 1983;28:116–9.
- 34 Melton CA, White BM, Davis RL, *et al*. Fine-tuned sentiment analysis of COVID-19 vaccine-related social media data: comparative study. *J Med Internet Res* 2022;24:e40408.
- 35 Health Map, Available: <https://www.healthmap.org/en/>

© 2023 Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.