BMJ Open

Developing an ethical framework-guided instrument for assessing bias in EHR-based Big Data studies: a research protocol

Shan Qiao ,[1] George Khushf,[2] Xiaoming Li,[3] Jiajia Zhang,[4] Bankole Olatosi [5]

To cite: Qiao S, Khushf G, Li X, *et al*. Developing an ethical framework-guided instrument for assessing bias in EHR-based Big Data studies: a research protocol. *BMJ Open* 2023;**13**:e070870. doi:10.1136/bmjopen-2022-070870

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-070870).

Received 07 December 2022
Accepted 03 July 2023

For numbered affiliations see end of article.

Correspondence to
Dr Bankole Olatosi;
olatosi@mailbox.sc.edu

ABSTRACT

Introduction The emergence of Big Data health research has exponentially advanced the fields of medicine and public health but has also faced many ethical challenges. One of most worrying but still under-researched aspects of the ethical issues is the risk of potential biases in data sets (eg, electronic health records (EHR) data) as well as in the data curation and acquisition cycles. This study aims to develop, refine and pilot test an ethical framework-guided instrument for assessing bias in Big Data research using EHR data sets.

Methods and analysis Ethical analysis and instrument development (ie, the EHR bias assessment guideline) will be implemented through an iterative process composed of literature/policy review, content analysis and interdisciplinary dialogues and discussion. The ethical framework and EHR bias assessment guideline will be iteratively refined and integrated with preliminary summaries of results in a way that informs subsequent research. We will engage data curators, end-user researchers, healthcare workers and patient representatives throughout all iterative cycles using various formats including in-depth interviews of key stakeholders, panel discussions and charrette workshops. The developed EHR bias assessment guideline will be pilot tested in an existing National Institutes of Health (NIH) funded Big Data HIV project (R01AI164947).

Ethics and dissemination The study was approved by Institutional Review Boards at the University of South Carolina (Pro00122501). Informed consent will be provided by the participants in the in-depth interviews. Study findings will be disseminated with key stakeholders, presented at relevant workshops and academic conferences, and published in peer-reviewed journals.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study will advance our understanding of bias and equity issues in Big Data research and develop an ethical framework and a guideline for assessing bias in electronic health records (EHR)-based Big Data studies.

⇒ This study will combine perspectives of both ethical study and data science and take advantage of integration of literature and qualitative data through the integrative process.

⇒ The developed ethical framework and EHR bias assessment guideline will be pilot-tested within an ongoing EHR project.

⇒ The in-depth interviews will be conducted among the key stakeholders in South Carolina, but this may not reflect the full range of insights from researchers and key stakeholders engaged in Big Data health research in other social contexts.

INTRODUCTION

The emergence of Big Data health research, characterised by tremendously large electronic health records (EHR) data sets and computational technologies such as artificial intelligence (AI) and machine learning (ML),[1] has exponentially advanced the fields of medicine and public health by making possible a better understanding regarding social determinants of health; discovering novel treatments; and mapping the underlying mechanisms, markers and progression of disease.[2–6] While widely used in diagnosis, clinical decision-making and personalised medicine, AI/ML, as a collection of data-driven technologies, has raised a novel set of ethical challenges, including respecting patient autonomy, adequate consent, identifiability and privacy protection and data ownership, sharing and reuse.[7–10] In alignment with the FAIR principles for scientific data (ie, Findable, Accessible, Interoperable, Reusable),[11] the usage of EHR data in biomedical and behavioural research should be guided by a sound ethical framework with steps taken to minimise unintended harm that could result from Big Data health research.[12–14]

Current policy and ethical guidelines for Big Data research, however, lag behind the technological progress being made in the healthcare field.[15–17] One of the ethical challenges encountered by Big Data research using EHR data is how to assess potential biases in its data curation, acquisition, linkage and integration.[18–21] For example, when data are predominately obtained from a single group, based on race/ethnicity, country of origin or socioeconomic status, the research can help over-represented populations, while not benefitting, and even potentially harming under-represented populations (group harm).[22–24] In addition to unrepresentative data, Big Data research may challenge equity through AI/ML algorithms trained using biased data (eg, data with a large number of missing/incomplete records).[25 26] The biased results can perpetuate existing health disparities and may even automate structural discrimination resulting in group harm.[27–29]

Despite increased concerns about potential biases in EHR data and data acquisition processes, very little Big Data research using EHR data reports biases in data or data acquisition and/or mining as an indicator of the research quality.[30] Such a limitation largely results from several knowledge gaps: (1) Lack of an ethical framework as a theoretical ground to study the bias in EHR data and/or necessary mitigation strategies; (2) lack of standardised measurement instruments or guideline to assess to what extent biases intentionally and/or accidentally emerge from the multiple steps of the Big Data curation cycle; and (3) lack of effective interdisciplinary collaboration that engages ethics experts, professional data curators, data management experts, data repository administrators, healthcare workers and state agencies in discussions addressing this ethical challenge.

To address the existing knowledge gaps in the ethical development of Big Data health research, the main goal of our study is to develop an ethical framework-guided instrument for assessing biases in EHR data with the following aims: (1) To develop an ethical framework for unbiased and inclusive Big Data research which will guide the development of an instrument in this study as well as future work in developing ethical principles and standards for Big Data health research; (2) to create and modify the instrument (ie, EHR bias assessment guideline) to assess potential biases in EHR studies; and (3) to pilot test the EHR bias assessment guideline for its applicability in an ongoing NIH-funded Big Data HIV project (R01AI164947; see online supplemental appendix 1 for a brief project description).

METHODS AND ANALYSIS

Conceptual framework of the study

The blueprint of our study can be presented by a conceptual framework (figure 1). Ethical analysis and EHR bias assessment guideline development will be implemented through an iterative process composed of literature/policy review, conceptual analysis and interdisciplinary dialogues and discussion. The ethical framework and EHR bias assessment guideline will be iteratively refined and integrated with preliminary summaries of results in a way that informs subsequent research. We will engage data curators, end-user researchers, healthcare workers and patient representatives throughout all iterative cycles using various formats including in-depth interviews of key stakeholders, panel discussions and charrette workshops. An initial conceptual analysis regarding bias issues in Big Data research based on literature/policy review will inform the ethical framework and EHR bias assessment guideline development. The rich evidence based on in-depth interviews, interdisciplinary dialogues and community charrette of diverse key stakeholders regarding realistic constraints and potential actions will be the pragmatic, reality stratum. The dialogues and integration of multiple iterative cycles will lead to refined versions of the ethical framework and the EHR bias assessment guideline. We will then pilot test and finalise this guideline in one ongoing Big Data project. The project is planned to be implemented from August 2022 to August 2023.

Literature/policy review

Literature/policy review will be conducted as a ground for developing the initial ethical framework and a metric tool. We plan to search at least six databases (PsycINFO, SocINDEX, PhilPapers, CINAHL, PubMed and Web

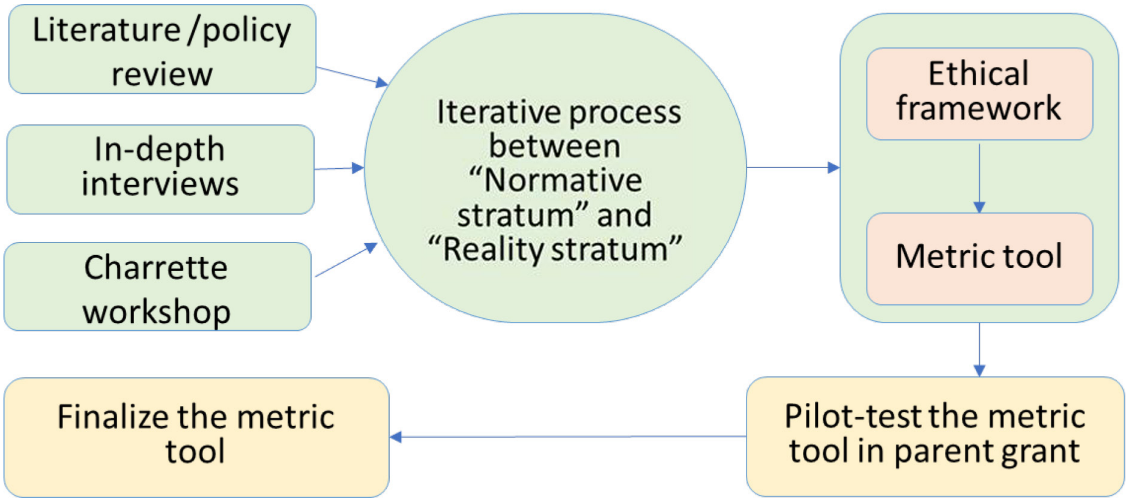
Literature /policy review
In-depth interviews
Charrette workshop
Iterative process between "Normative stratum" and "Reality stratum"
Ethical framework
Metric tool
Pilot-test the metric tool in parent grant
Finalize the metric tool

Figure 1 Conceptual framework of the study.

Table 1 Data resources of the policy review

Relevant laws and regulations	Relevant online ethical teaching resources
Gramm-Leach-Bliley Act	National Collaborating Centres.
Health Insurance Portability and Accountability Act	Resources for Research Ethics Education.
European Union General Data Protection Regulation	Ethics Committee of the American College of Epidemiology.
Framework for responsible sharing of genomic and health-related data by Global Alliance for Genomics and Health	Canadian Tri-Council Policy Statements on the Ethic for Research Involving Humans.

of Science) using search terms such as ‘Big Data’, ‘data mining’, ‘algorithms’, ‘bias’, ‘ethic*’, ‘electronic *record’, ‘inclusive’, ‘equity’, ‘equality’ and ‘*justice’. We will also include common qualifiers for health inequalities, such as gender, race, ethnicity, socioeconomic status and stigma to produce relevant search results because they are more specific when identifying sources of bias. These terms will be combined using Boolean logic. The inclusion criteria will be (1) papers published in English; (2) papers related to Big Data research; and (3) papers focused on ‘bias’ or ‘equity’ issues. That is, we will include all the studies about biases in EHR studies using the Big Data approach, regardless of whether or not the studies explicitly investigate the relationship between bias and equity. To obtain a broader understanding of this ethical challenge related to Big Data research, no restrictions will be placed on the discipline of the papers or on the type of methodology. To examine current legal/ethical frameworks and principles used in guiding data-driven research, especially EHR-based studies, we will search and review relevant guidance, laws and regulations[7] (see data sources in table 1).

All selected literature/documents will be reviewed using the thematic synthesis method with three steps: coding of text ‘line-by-line’, development of ‘descriptive themes’ and the generation of ‘analytical themes’.[31] Research assistants with training in ethics and experience in Big Data research will conduct literature/policy review under supervision. Any disagreement about paper screening and information synthesising will be resolved by team discussion and decided by principal investigator (PI).

Conceptual analysis

Grounded in the results of the literature/policy review, we will conduct a conceptual analysis with aims of clarifying concepts of ‘bias’ and ‘equity’ associated with Big Data research so we can develop an ethical framework to guide identifying and measuring bias. Rodgers’s well-established method of conceptual analysis will be employed, which presents an inductive, dynamic view of a phenomenon.[32] Data extracted from the relevant literature will be categorised as: (1) defining attributes (characteristics of the concept); (2) antecedents; and (3) consequences. Verbatim statements from each article will be tabulated. An inductive analysis of ‘bias’ and ‘equity’ will produce descriptive themes. We will identify what types of biases in EHR data curation, acquisition and process are key issues from the ethical perspectives. These biases will be a focus of the EHR bias assessment guideline to be developed.

In-depth interviews

Since key stakeholders may have diverse opinions and perspectives about the ‘unbiased’ Big Data study, we believe that in-depth interviews will be an appropriate approach to collecting qualitative data, which will offer opportunities for one-on-one, in-depth conversations with minimum influence of others on the interviewee.

The key stakeholders of the ongoing Big Data project include but are not limited to: (1) Two key state partners: the South Carolina Department of Health and Environmental Control (SC DHEC) and the South Carolina Revenue and Fiscals Affairs Office (SC RFA), who have been actively involved in our research as partners since 2017 through multiple EHR-based Big Data research projects; and (2) a functional Stakeholder Advisory Board (SAB), which includes five to seven members representing the relevant stakeholders (eg, SC DHEC, HIV treatment and care physicians and People living with HIV/AIDS (PLWH).

We will purposely recruit about 20 participants among the key stakeholders for the interviews including data scientists and research staff (n=5), healthcare workers in HIV clinics (n=5), representatives of relevant state agencies (n=5) and patient representatives (ie, people living with HIV) (n=5) in SC. The research team will contact and recruit the participants based on recommendations by the SAB. Written consent will be obtained prior to the interviews. A semi-structured qualitative interview guide will be created by researchers who have extensive experience working with key stakeholders. The questions will be tailored for different types of participants but will generally focus on perceptions and understandings of bias and equity in the context of Big Data research, the criteria of an unbiased and quality Big Data study, potential causes of bias in EHR-based studies, the challenges and barriers to conducting unbiased Big Data studies and the possible solutions or policies for addressing these problems. Additional topics will be included as appropriate and as informed by the ethical framework. With appropriate consent, the interviews will be audio-recorded. Interviewers will take field notes during the interviews to serve as a complementary data source. Each interview will take 1 hour led by trained interviewers in a privacy room or via online conferencing per request.

All interviews will be transcribed verbatim and entered into software NVivo by research staff. Preliminary coding will begin by reading and rereading five transcripts. A codebook will then be developed to include both deductive (ie, the themes drawn from the conceptual framework) and inductive (ie, the new themes emerging from the interviews) codes. Two research staff will independently code each of the transcripts using the codebook. Any coding disagreements will be discussed and resolved using a consensus model of team-based coding. Quote excerpts and coding memos will be developed according to themes. Representative and verbatim quotes will be selected to illustrate key findings.[33 34]

Interdisciplinary dialogue through charrette workshops

Charrette workshops with scholars and key stakeholders will be organised to promote interdisciplinary dialogues and discussions about the framing of ethical issues and the development of the EHR bias assessment guideline.

As a community engagement strategy recommended by the National Minority AIDS Council, a charrette is a collaborative planning process that purposefully brings together the expertise of community and academic research partners to strengthen partnerships, engage stakeholders and make decisions regarding translational research.[35] We will invite ethics experts, professional data curators, data management experts, data repository administrators, healthcare workers, representatives of relevant state agencies and PLWH through the SAB of the ongoing Big Data project. To obtain a wider healthcare perspective, we will also invite experts in other health conditions and/or from a broader academic network of Big Data health studies leveraging the Big Data Health Science Center in our institute. We will assemble a panel of 10–15 experts for a 1-day workshop to discuss the draft EHR bias assessment guideline. Panellists will receive the guideline draft 2 weeks prior to the charrette and be asked to review and provide feedback on its content, structure and format. The charrette will be held in a University of South Carolina conference room or conducted in a Zoom platform using the 'breakout discussion room' function, depending on the logistics. The charrette will begin with a review of the charrette goals and an explanation of the procedures for the day. It will be highlighted at the very beginning of the workshop that there are no 'right' or 'wrong' answers and that principles of respect and openness during the dialogues create a safe and comfortable environment for discussion. Panellists will be divided into groups of 3–4. The research team will co-lead each of the small group discussions. Each group will discuss the same sets of questions that are based on the charrette objective (eg, feedback on the tool, strengths and weakness, additional content), and a co-leader will record the primary points on a discussion board. After completing the small group discussions, the full group will reconvene and a representative from each group will present their findings; other members will ask questions for points of clarification, and additional information will be added to the discussion board if needed. The discussion notes will become the primary data source. Field notes will be taken during the charrette by two research staff, with observational and interpretive elements. At the end of the charrette, the research team will engage in a process of critical reflection regarding the group and develop combined reflection notes based on these conversations.

Ethical framework and guideline development: an integrative process

The development of the ethical framework and EHR bias assessment guideline will be an integrative process informed by all the knowledge and qualitative research obtained through literature/policy review, in-depth interviews and charrette workshops. Specifically, the literature/policy reviews will advance our understanding of the landscape of ethical development and relevant topics and debates about using EHR data in healthcare research. The interdisciplinary communication and discussions among key stakeholders will further help us to identify the key types of biases in EHR-based studies that have ethical implications in core values such as 'social bias', 'equity' and 'justice'. The initial ethical framework and EHR bias assessment guideline will be refined based on multiple iterative cycles in which preliminary summaries of results (based on literature/policy review and qualitative studies) inform research in subsequent steps until the research team believe that the refined version of ethical framework and the EHR bias assessment guideline comprehensively reflects and integrates key issues based on both 'normative stratum' and 'reality' stratum.

Grounded in the reviews and the ethical framework, we could focus on the 'ideal', normative stratum of the EHR bias assessment guideline, that is, 'what should be'. The rich qualitative evidence from the key stakeholders among the front-line data curators, healthcare providers and end-user researchers and patients can be used to build up the 'reality' stratum of the EHR bias assessment guideline, that is, 'what realistic constraints' and 'what could be', to ensure that this assessment guideline is applicable and reasonable in real-world practice. The lived experiences, reflection and lessons from data curators, management experts and repository administrators will assist us in criteria/standards selection and adaptation. The research team will actively participate in the discussions to integrate the two strata of the metric tool and ultimately develop the EHR bias assessment guideline, that is, informed by the ethical framework and also rooted in the realities of EHR data curation, acquisition, process and usage in health fields.

Extant literature suggests that strategies for resolving potential biases in EHR studies are context dependent. It is not unusual that one approach of addressing/adjusting for one type of bias may cause another type of bias due to the complexities of the clinical data set and the healthcare system. Therefore, the EHR bias assessment guideline will not aim to provide comprehensive resolving approaches or methods or cover all types of biases. Rather, it is more like a checklist of potential key biases in EHR studies from data acquisition, data integration and data process. It focuses on key bias issues from the ethical perspective (eg, related to equity), and will follow the format of several widely accepted assessment instruments of the quality of clinical trials.[36–39] Scores will reflect the level of concern that the researchers identify and assess the ethically important biases in key steps of the data curation, acquisition and analysis cycle. Therefore, this guideline will assist Big Data researchers in knowing, assessing, reporting and resolving potential biases in EHR data.

Pilot testing the EHR bias assessment guideline

The pilot test using the EHR data in the existing Big Data project will focus on (1) assessing if the criteria/standards of the EHR bias assessment guideline are applicable and reasonable in the specific setting and scenario of the project; (2) adapting and refining the content and format of the EHR bias assessment guideline to ensure that this instrument is valid and reliable when used in a real-world practice setting and the language is precise, accurate and easy to follow; and (3) identifying any additional criteria or items that are needed for the existing instrument based on complex data curation of the project. The pilot test will be based on data sources in the existing project, including study protocols and all of the relevant documents (eg, data dictionaries, contracts with the SC RFA) of the ongoing Big Data project; meeting records of quarterly SAB meetings since the beginning of the project; minutes of research team meetings of the Big Data project; and preliminary data from the ongoing Big Data project if available. Through reviewing the research documents of the ongoing Big Data project, we will assess the potential bias in EHR data curation and processing using the EHR bias assessment guideline. We will discuss the findings with the research team and SAB to contextualise the findings (eg, the level of bias) in the real-world settings of acquiring and using EHR data in the ongoing Big Data project to address research questions on viral suppression. Finally, we will hold multiple group discussions with the research team and stakeholders of the ongoing Big Data project, key informants who participated in previous phases of this study (eg, in-depth interview), as well as external experts to go through the EHR bias assessment guideline and findings and collect feedback and suggestions for refinement. We will use an iterative process with interactive strategies, whereby notes of discussions taken during each meeting will be triangulated with other notes of document reviews in finalising the EHR bias assessment guideline.

PATIENT AND PUBLIC INVOLVEMENT

Key stakeholders of the proposed study will be involved in study design, conduct and reporting of our research. We will actively reach out to patients and public in the dissemination of our findings.

ETHICS AND DISSEMINATION

The study was approved by the nstitutional Review Board at the University of South Carolina (Pro00122501). Informed consent will be provided by the participants in the in-depth interviews. Study findings will be disseminated with key stakeholders, presented at relevant workshops, academic conferences and published in peer-reviewed journals.

DISCUSSION

Our study has several strengths. First, the EHR bias assessment guideline will be informed by an ethical framework. An assessment guideline informed by an ethical framework will integrate ethical principles and technical realities and thus promote both ethical development and EHR data application in healthcare fields. The dialogues and communications between ethics and data science will increase the awareness of ethical challenges among key stakeholders. Second, we have an interdisciplinary team that includes ethics and data science experts, social scientists, state-wide data repository managers and HIV experts and clinicians with a proven history of working collaboratively in publication and grant application. This team will be able to comprehensively understand the bias issue rather than isolate questions of bias 'in' data sets. Third, we will apply sequential use of multimethod data collection (literature review, in-depth interviews, community charrettes via a workshop) and analysis strategies in a participatory manner, which allows for unique mixed-methods findings. The engagement of diverse key stakeholders in the data collection will assure that multiple voices from various communities (healthcare providers, healthcare agencies, government, patients and researchers) will be incorporated and given priority. Fourth, we will also invite external experts and key informants in the EHR bias assessment guideline development and pilot testing to obtain a wider perspective of public health beyond the HIV-related project.

However, our study has several limitations. First, in-depth interviews will be conducted among the key stakeholders in South Carolina and this may not reflect the insights of researchers and key stakeholders engaged in Big Data health research in other social contexts. Second, although we try to broaden our perspectives by engaging experts and key stakeholders in various public health areas, the pilot test will be applied in an ongoing HIV project. Therefore, future studies may need to further refine and modify the ethnical framework and EHR bias assessment guideline so they can be adopted to more disease conditions.

Despite these limitations, the outputs of our study will advance our understanding of bias and equity issues in Big Data research, develop an ethical framework and an EHR bias assessment guideline for assessing bias in EHR-based Big Data studies, and thus lead to and inform a more nuanced assessment and exploration of bias in practice for ethical development of Big Data health research beyond the existing Big Data project. The guideline can be reused as an assessment instrument to detect and quantify bias, which may contribute to improving awareness and exploration of this critical ethical challenge. The ethical framework may also provide insights and guidance for addressing bias issues in Big Data using other types of data beyond EHR.

Author affiliations

[1]Health Promotion Education and Behavior, University of South Carolina, Columbia, South Carolina, USA
[2]Department of Philosophy, University of South Carolina, Columbia, South Carolina, USA
[3]Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA
[4]Department of Epidemiology and Biostatistics, Arnold School of Public Health, South Carolina College of Pharmacy - University of South Carolina Campus, Columbia, South Carolina, USA
[5]Health Services, Policy and Management, University of South Carolina Arnold School of Public Health, Columbia, South Carolina, USA

Acknowledgements The authors thank SC Department of Health and Environmental Control, SC Revenue and Fiscal Affairs Office and other SC agencies as key stakeholders of the proposed study. The project is funded by National Institute of Allergy and Infectious Diseases (NIH/NIAD) under award number R01AI164947.

Contributors SQ, XL, BO and GK conceptualised the study and SQ wrote the first draft. XL, JZ and GK participated in writing sections of the original proposal. All authors critically reviewed and edited the manuscript. BO completed the Institutional Review Board approvals. BO and JZ secured the funding.

Funding Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (NIH/NIAD) under award number R01AI164947.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

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.

ORCID iDs

Shan Qiao http://orcid.org/0000-0003-1834-1834
Bankole Olatosi http://orcid.org/0000-0002-8295-8735

REFERENCES

1 Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. *NPJ Digit Med* 2018;1:53.
2 Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the Hype? *JAMA* 2019;321:2281–2.
3 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;546:686.
4 Ahmed Z, Mohamed K, Zeeshan S, *et al*. Artificial intelligence with multi-functional machine learning platform development for better Healthcare and precision medicine. *Database (Oxford)* 2020;2020:baaa010.
5 Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017;69S:S36–40.
6 Luxton DD. An introduction to artificial intelligence in behavioral and mental health care. In: *Artificial intelligence in behavioral and mental health Care*. Elsevier, 2016: 1–26.
7 Salerno J, Knoppers BM, Lee LM, *et al*. Ethics, big data and computing in epidemiology and public health. *Ann Epidemiol* 2017;27:297–301.
8 Bourne PE. Confronting the ethical challenges of big data in public health. *PLoS Comput Biol* 2015;11:e1004073.
9 Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018;378:981–3.
10 Murphy K, Di Ruggiero E, Upshur R, *et al*. Artificial intelligence for good health: a Scoping review of the ethics literature. *BMC Med Ethics* 2021;22:14.
11 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
12 Jacobsen A, De Miranda Azevedo R, Juty N, *et al*. *FAIR principles: interpretations and implementation considerations*. Cambridge: MIT Press One Rogers Street, 2020: 10–29.
13 Ballantyne A. Adjusting the focus: a public health ethics approach to data research. *Bioethics* 2019;33:357–66.
14 Dignum V. *Ethics in artificial intelligence: introduction to the special issue*. Springer, 2018: 1–3.
15 Gray EA, Thorpe JH. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. *J Comp Eff Res* 2015;4:61–74.
16 Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;25:37–43.
17 Ferretti A, Ienca M, Hurst S, *et al*. Big data, BIOMEDICAL research, and ethics review: new challenges for Irbs. *Ethics &Amp; Human Research* 2020;42:17–28. 10.1002/eahr.500065 Available: https://onlinelibrary.wiley.com/toc/25782363/42/5
18 Crawford K. The hidden biases in big data. *Harv Bus Rev* 2013;1.
19 Saka E. n.d. Big data and gender-biased Algorithms. *The International Encyclopedia of Gender, Media, and Communication*;2020:1–4.
20 Howe III EG, Elenberg F. Ethical challenges posed by big data. *Innov Clin Neurosci* 2020;17:24.
21 Brodie MA, Pliner EM, Ho A, *et al*. Big data vs accurate data in health research: large-scale physical activity monitoring, Smartphones, Wearable devices and risk of unconscious bias. *Med Hypotheses* 2018;119:32–6.
22 Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. *JAMA Intern Med* 2019;179:293–4.
23 Lewis CM, Obregón-Tito A, Tito RY, *et al*. The human Microbiome project: lessons from human Genomics. *Trends Microbiol* 2012;20:1–4.
24 Angwin J, Larson J, Mattu S, *et al*. Machine bias. In: *Ethics of Data and Analytics*. Auerbach Publications, 2016: 254–64.
25 Obermeyer Z, Mullainathan S. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. Proceedings of the conference on fairness, accountability, and transparency; 2019
26 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the Healthcare system: retrospective observational study. *BMJ* 2018;361:k1479.
27 Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Sci Eng Ethics* 2018;24:1521–36.
28 Lipworth W, Mason PH, Kerridge I, *et al*. Ethics and epistemology in big data research. *J Bioeth Inq* 2017;14:489–500.
29 Crawford K, Gray ML, Miltner K. Big Data| Critiquing big data: politics, ethics, Epistemology| special section introduction. *International Journal of Communication* 2014;8:10.
30 Favaretto M, De Clercq E, Elger BS. Big data and discrimination: perils, promises and solutions. A systematic review. *J Big Data* 2019;6:1–27.
31 Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol* 2008;8:45.

32 Rodgers BL. Concepts, analysis and the development of nursing knowledge: the evolutionary cycle. *J Adv Nurs* 1989;14:330–5.

33 Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing Taxonomy, themes, and theory. *Health Serv Res* 2007;42:1758–72.

34 Pope C, Ziebland S, Mays N. Analysing qualitative data. *Bmj* 2000;320:114–6.

35 Samuel CA, Lightfoot AF, Schaal J, *et al*. Establishing new community-based Participatory research partnerships using the community-based Participatory research Charrette model: lessons from the cancer health accountability for managing pain and symptoms study. *Prog Community Health Partnersh* 2018;12:89–99.

36 Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary *Control Clin Trials* 1996;17:1–12.

37 van Tulder M, Furlan A, Bombardier C, *et al*. Updated method guidelines for systematic reviews in the Cochrane collaboration back review group. *Spine (Phila Pa 1976)* 2003;28:1290–9.

38 Higgins JPT, Thomas J, Chandler J, *et al*. Cochrane Handbook for systematic reviews of interventions. In: *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 23 September 2019.

39 Kim KS, Jo JK, Chung JH, *et al*. Quality analysis of randomized controlled trials in the International Journal of impotence research: quality assessment and relevant clinical impact. *Int J Impot Res* 2017;29:65–9.

Appendix 1 Introduction of the ongoing Big Data project

Summary of the ongoing Big Data project (R01AI164947). Applying Big Data analytics to integrated EHR datasets of people living with HIV (PLWH) in South Carolina (SC), the ongoing Big Data project proposed to examine the longitudinal dynamic pattern of viral suppression (VL), develop optimal predictive models of various viral suppression indicators, and translate the models to service-ready tools for clinical use. The ongoing project (R01AI164947) will extract longitudinal EHR and other relevant data of all PLWH in SC from 8 state agencies and then link the patient-level data with county-level data (e.g., healthcare infrastructure indicators) from multiple publicly available data sources (2005 to 2020) to address the following specific aims: Aim 1: Identify the longitudinal dynamics of VL among PLWH in SC using multiple indicators, including time to initial suppression, sustained suppression, viral rebound, viral blips, and other relevant VL measures. Aim 2: Determine the critical predictors of multiple VL indicators through AI-based modeling accounting for factors at the individual level, structural level, and socioenvironmental level. Aim 3: Develop a multifactorial decision system based on a risk prediction model to assist with identifying the risk of viral failure or viral rebound when patients present at clinical visits.

Data source and structure. Eight state agencies/systems that oversee and administer public data repositories in SC have contributed EHR and other relevant data to the existing BIg Data project. They are SC Department of Health and Environmental Control (SC DHEC) Enhanced HIV/AIDS Reporting System (eHARS), Ryan White HIV/AIDS Program Data Report (RDR), SC Revenue and Fiscal Affairs Office (SC RFA), Health Sciences South Carolina (HSSC), SC Department of Mental Health (DMH), SC State Law Enforcement Division (SLED), SC Department of Alcohol and Other Drug Abuse Services (DAODAS), and Prisma. SC RFA serves as the honest broker for the linkage of all identifiable data from various sources and removes all identifiable information from the linked data before releasing it to the research team, who enters into legal contract (s) (required for each new study or new analysis) with SC RFA.