Big Data, Biomedical Research, and Ethics Review: *New Challenges for IRBs*

Agata Ferretti, Marcello Ienca, Samia Hurst, and Effy Vayena

ABSTRACT The increased use of big data in the medical field has shifted the way in which biomedical research is designed and carried out. The novelty of techniques and methods brought by big data research brings new challenges to institutional review boards (IRBs). Yet it is unclear if IRBs should be the responsible oversight bodies for big data research and, if so, which criteria they should use. A large but heterogenous set of ethics guidelines and normative responses have emerged to address these issues. In this study, we conducted a scoping review of soft-law documents and guidelines with the aim of assessing ongoing normative efforts that are proliferating in this domain. We also synthesize a set of recurrent guidelines that could work as a baseline to create a harmonized process for big data research ethics.
KEYWORDS big data, big data research, biomedical research, research ethics, institutional review board (IRB)
Ferretti, A., et al., "Big Data, Biomedical Research, and Ethics Review: New Challenges for IRBs," *Ethics & Human Research* 42, no. 5 (2020): 17-28. DOI: 10.1002/eahr.500065

Traditionally, human subjects research in the biomedical field engages healthy and sick people as research participants in order to test certain hypotheses about health and disease according to a well-defined study design. The study designs, such as randomized controlled trials and cohort studies, are carefully reviewed by institutional review boards (IRBs)—also known as *ethical review committees* (*ERCs*) in some countries. IRBs are committed to protecting the rights and welfare of human subjects recruited to participate in biomedical or behavioral research (including social science research).[1] This approach has long been the standard for biomedical research.

Recently, however, biomedical research has begun to pursue opportunities afforded by big data. Big data research relies on large-scale databases, multiplication of data sources, advanced storage capacity, and novel computational tools that allow for high-velocity data analytics.[2] In the biomedical domain, big data trends are enabled by and allow for advances in areas such as whole genome sequencing, brain imaging, mobile health, and digital phenotyping.[3] Today, a large portion of health-related research relies on big data. Big data also enables researchers to draw health insights from data sources that are not strictly medical—data from wearable trackers, social media, and Internet searches, for example.[4] Big data research opens new prospects to accelerate health-related research and potentially elicit breakthroughs that will benefit patients.[5]

Big data has been observed to shift the way biomedical researchers design and carry out their studies.[6] This research departs from the traditional research model because it is largely exploratory rather than hypothesis driven. Health-related big data research is based on the acquisition of large amounts of data from multiple and

often heterogeneous sources, which are subsequently combined and mined using powerful data analytics tools. This reverse-engineered approach to health-related research allows researchers to extract features and valuable insights from large datasets, without being able to anticipate exactly what the data analysis will find.

The methodological novelty of big data research models brings new challenges and questions to IRBs, including whether they are the bodies responsible for assessing these projects, and if they are, what criteria they should use to evaluate them. Given current technologies, analytic methods, and regulations, IRBs cannot take their traditional review frameworks as given. This is because big data research models might not fit within the traditional national review policies for the protection of human subjects (for example, the Common Rule in the United States and the Human Research Act in Switzerland) and the principles stated in guidelines documents such as the Helsinki Declaration of the World Medical Association and the U.S. National Commission's *Belmont Report*. It was observed that the definition of "human subjects" in the Common Rule might not cover big data projects involving the processing of deidentified data.[7] The Common Rule's scope, in fact, is limited to the acquisition and processing of "identifiable private information." As a consequence, privately held, publicly available datasets such as Twitter data might be considered exempt from IRB oversight, even though it is possible to reidentify those data sources by matching them with ancillary information.[8] Similarly, the European Union's research ethics legislation,[9] as well as the Swiss Human Research Act,[10] might not apply to research that involves anonymized data or secondary use of data for which a broad consent and an ERC approval was obtained. For instance, in Denmark, researchers can reuse genomic data previously extracted from a donated tissue sample of the National Biobank Registry for a new project without seeking ERC approval.[11]

Health-related big data research also challenges IRBs in referring to existing safeguards for ethics research such as informed consent, privacy and confidentiality, and minimal risk.[12] The reason for that stems from a threefold consideration.

First, individuals whose data are used in research (hereafter data subjects) are often not sufficiently informed concerning the use of their data. Particularly, researchers might not be able to adequately inform data subjects when collecting their data stored in large repositories or when mining the data in the context of secondary data uses. In a more pragmatic sense, informed consent might be hard to obtain in big data studies due to the high number of data subjects involved. This is particularly true when consent is sought retrospectively. In cases where research is conducted on large-scale repositories, it might be hardly feasible to recontact all data subjects and inform them that the purpose of data processing has changed from the original consent agreement stipulated at the time when the repository was created.

Second, breaches in data privacy and confidentiality represent a major source of risk for health research using big data. The reason for that stems from the informational richness of large research data repositories, which makes them a primary target for actors outside the research domain. Insurers, marketing companies, and the government might require access to these data. Furthermore, health-related data repositories have often been exposed to illicit use by malevolent actors. Although these repositories are usually composed of data that do not contain personal identifiers (deidentified data), research has shown that both pseudonymized data (with which artificial identifiers are used so data subjects can be reidentified) and anonymized data (from which actual and artificial identifiers are excluded in an effort to make reidentification impossible) could be matched with publicly available information or auxiliary data to allow the reidentification of a subject.[13] This reidentification risk is particularly problematic for health research, as health data constitute a highly sensitive data source.

Finally, correlations arising from health-related big data analytics can be abused by various actors for unethical purposes such as discriminating against applicants to health insurance services or jobs based on health risk indicators. These indicators include, among others, risk factors associated with genetic variants, neuroimaging biomarkers of addiction or antisocial behavior, and molecular biomarkers of chronic illness. This risk of discrimination also applies to not strictly medical data such as online behavioral information. For example, a recent study has used a big data approach to predict people´s sexual orientation from their online

behavior.[14] This poses a risk to many people, especially in countries where nonheterosexual behavior is prohibited by law. Although the research involved the processing of seemingly innocuous data points, the findings suggest that the risk of reidentification is potentially greater than minimal risk.

Many ethical issues remain to be solved. These include whether and when big data projects using deidentified data from public databases should require IRB approval, what counts as "public data," what constitutes "minimal-risk" in data-driven projects, and which novel ethical safeguards, if any, are required to ensure ethical big data research.

To reduce this uncertainty, various stakeholders have issued nonbinding guidelines. The scientific community has developed best-practice guidelines and educational activities aimed at sensitizing researchers about the ethical promises and challenges of big data research. In parallel, a growing number of professional organizations are restructuring their codes of conduct and providing research ethics training to data and computer scientists. Members of the scientific community, such as the editors of the journal *Nature*, have encouraged policy-makers to "further support such efforts . . . and make them better known to researchers" and have proclaimed that "all researchers have a duty to consider the ethics of their work beyond the strict limits of law or today's regulations."[15]

Nevertheless, the proliferation of many independent responses around big data research ethics has generated uncertainty among IRBs. This fragmented landscape of responses increases the confusion and leaves IRBs with unclear normative guidance about how to tackle big data ethics issues. We monitored and evaluated these efforts to bring clarity about the plurality of perspectives emerging in this domain. To accomplish our purpose, we have conducted a scoping review of the soft-law documents and guidelines[16] concerning the ethics of health-related big data research. While previous reviews have screened the scholarly literature on this topic,[17] and opinion articles have discussed their implications for IRBs,[18] no other study, to our knowledge, has provided a comprehensive assessment of the emerging body of guidelines on this topic. Research best practices, recommendations, codes of conduct, and other guidance documents—especially those commissioned by funding agencies, professional associations, academic societies, nongovernmental organizations (NGOs), think tanks, and private companies—are typically not formally published in peer-reviewed academic journals but, rather, released in commissioned technical reports, white papers, and similar documents. Because these types of documents are usually not indexed in academic archives and databases, reviewing this gray literature[19] is critical to retrieve and assess this body of information. We believe that IRBs will benefit from our research, as we provide a comprehensive set of recommendations that could represent the starting point for IRBs in revising and harmonizing their ethics research processes.

A few documents addressed the issue of whether IRBs should be the oversight body accountable for big data research at all. Any development of an ethical framework for big data research, however, cannot disregard the active involvement of IRBs in decision-making.

APPROACH TO IDENTIFYING DOCUMENTS FOR REVIEW

In February 2019, we conducted an online search of the gray literature addressing the ethical implications of health-related big data research. Gray literature is defined as "literature that is not formally published in sources such as books or journal articles."[20] This definition includes nonconventional material such as reports, technical specifications and standards, technical and commercial documentation, official documents,[21] and "that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishing interests and where publishing

is not the primary activity of the organisation."[22] Gray literature reviews have been observed to have the threefold advantage of providing information of process and implementation, both of which can be missing from scientific papers;[23] reducing the typical publication lag of peer-reviewed articles, hence ensuring more efficient responses;[24] and validating the results of research-based literature searches.[25]

Following previous studies,[26] we used a multistage screening process involving both inductive screening via search engine and deductive identification of relevant agencies (for example, national, international, and intergovernmental organizations), and subsequently we screened their websites and online collections. A total of 49 documents were included in our analysis (see appendix 1, available online, along with all the figures and the appendices; see "Supporting Information" at the end of this article). In the literature retrieval phase, we searched the Google engine in nonpersonalized mode using multiple combinations of the following keywords: "big data," "data science," "digital data," "medical," "healthcare," "clinical," "policy," "ethics," "governance," "ethics committee," "IRB," and "ethics review board." Combinations of keywords were reiterated until saturation was achieved. In the screening phase, we selected, on the one hand, soft-law documents issued by national and international agencies (highlighted in blue in appendix 1) and, on the other hand, nonlegal guidelines providing best practices and recommendations, disseminated by NGOs, professional organizations, research bodies, private companies, think tanks, and other actors (highlighted in yellow in appendix 1).[27] We included only official documents representative of and issued by collective entities. Documents such as personal blogs, written and issued by individual authors offering their personal views, were not included. In the filtering phase, we excluded from the analysis all documents that did not meet our content-based inclusion criteria (see appendix 2). The documents were collected independently by two coauthors who compared their results and resolved interpretative discrepancies upon discussion. Documents written in English, Italian, French, Greek, and German (languages spoken by research team members) were included in the analysis. We then conducted a descriptive numerical summary and a thematic analysis. The descriptive numerical summary consisted of calculating the frequencies of the total number of articles included, the distribution of documents by the documents' issuers and the targeted stakeholder group, and the prevalence of documents with a particular health-related or IRB focus. In the latter analysis, two researchers inductively identified recurrent themes[28] with software assistance. The two researchers coded the themes using the NVivo software for qualitative data analysis (version 12 for Mac) considering three macro areas: general ethical issues, normative ethical recommendations, and specific recommendations for IRBs. Within these areas, we grouped and merged our coding in themes and macrothemes.[29] Disagreement about where to allocate codes that did not seem to follow in any existing theme was resolved among coauthors through internal consultation.

LITERATURE REVIEW RESULTS

Most documents were issued by national and governmental institutions (28%), followed by professional organizations (22%) and NGOs (20%). Fewer documents (no more than five) were issued by private companies, international institutions, research institutions, and think-tank platforms. The geographical provenance of the issuers (41% from North America, 35% from Europe, 20% being international ("international" meaning that the issuer of that document was not located in a specific country; rather, it was a multicountry agency), and 4% from Oceania) showed a higher representation of highly industrialized Western countries and a relative underrepresentation of low- and middle-income countries from the global south. The majority of the documents (35%) targeted readers from various stakeholder groups (for instance, government institutions, researchers, professionals, industry associations, consumer advocates, and the general public), with a quarter specifically targeting governmental regulations and a fifth addressing professional groups (e.g., the professional organization of statisticians) (see figure 1).

As to the content of the documents, 55% had a prominent focus on health-related big data, with the remaining 45% having a broader focus on big data research, including health-research themes. Additional analysis revealed that only 16 documents (33%) were addressed explicitly to IRBs or ERCs and provided ad

hoc recommendations for the review of big data projects. The remaining documents provided general recommendations concerning the general ELSI (ethical, legal, and social implications) of big data research, but they do not address IRBs directly.

Our inductive thematic analysis identified a number of mutually interconnected ethical themes (see figure 2). The notion of privacy was by far the most prevalent; it was mentioned in all documents and discussed in depth in three quarters of them. A highly recurrent issue concerned how to balance data providers' privacy while enabling the progress of research using big data techniques. While the importance of preserving privacy was widely recognized across the literature we reviewed, some documents raised the problem of harmonizing privacy regulations: given that privacy regulations differ across different jurisdictions, it might not, documents observed, be straightforward for researchers and IRBs to determine how to ensure privacy in cross-national big data research. Documents also addressed the problem of ensuring the semantic unambiguity of the privacy concept in spite of the blurred distinction between public and private data repositories in the digital ecosystem. For example, documents questioned whether health-related big data projects that use data from public-by-default social media platforms such as Twitter should undergo similar privacy impact assessments and ethics review as conventional biomedical research does.

The second most common ethical theme was informed consent (discussed in 44% of documents), whose ethical sensitivity was primarily associated with the problem of obtaining retrospective consent from data subjects when conducting large-scale big data studies. Documents questioned the practical feasibility of retrospectively contacting many hundreds of thousands of data subjects—such as during the emotional "contagion" study[30]—and discussed the ethical justification for seeking retrospective consent in the context of public health research conducted in the public interest (such as epidemic prevention). Data ethics issues associated with data management, such as data security (39%), data sharing (37%), and data transparency (40%), also composed a significant portion of the current ethical landscape. Issues of algorithmic bias, beneficence, the right to be forgotten, data ownership, and individual autonomy appeared less prevalent.

Contextual analysis of emergent themes identified the ethical, legal, social, or technical contexts within which each subtheme was discussed or a solution was proposed (see figure 2). Results showed that the documents discussed 41% of the themes in the context of normative ethical analysis and with respect to potential solutions relying on ethical guidelines, best practices, or conceptual clarification. A third of the themes (31%) were discussed with reference to their technical implications and solutions. For example, distributed-ledger computing (or blockchain), encryption, and differential privacy were all mentioned as possible technical solutions to privacy risks in the big data domain.[31] A smaller number of documents addressed the political and regulatory domain and proposed solutions in terms of novel legislation (18%) or social strategies (12%).

Analysis of thematic interrelations indicates a high degree of interconnectedness among (sub)themes and contexts. Although primarily discussed in the context of ethical analysis, privacy and informed consent issues were largely cross-discussed in various contexts including the technical and legislative domains. Issues of data security and stakeholder collaboration were primarily discussed in connection with, respectively, technical solutions and social considerations. Data security was primarily presented as a problem requiring technical solutions (for example, enhanced encryption, immunization to abusive apps, and adherence to international standards such as the ISO/IEC 27001:2013[32]), whereas hard law was considered unavoidable to address issues of data ownership and to enforce the data subjects' right to be forgotten. Among these regulatory solutions, the European Union's General Data Protection Regulation, implemented in May 2018, was perceived as setting a new standard for data subject rights. Our thematic analysis also retrieved substantive recommendations for the ethics review of big data research. These recommendations concerned a variety of thematic families (see figure 3) and presented notable divergences in terms of content and degree of specification.

From the perspective of substantive ethical content, the most recurrent recommendation was for IRBs to ensure that researchers are providing adequate information to data subjects as part of the process of obtaining their informed consent (n = 31, 63%). Further recommendations required data controllers and processers to

protect the privacy of data subjects (n = 30, 61%) and to assure data transparency and the trustworthiness of data-driven inferences (n = 26, 53%). A smaller portion of recommendations required oversight bodies and regulators to foster collaborative exchange among stakeholders about data uses (n = 23, 47%) and to provide new guidance to assess the benefits and harms of big data research (n = 21, 42%). Appendix 3 provides further information concerning the substantive ethical recommendations issued in different continents and by various stakeholders.

In terms of granularity or degree of specification, most recommendations made general normative statements about the importance of promoting certain ethical principles (such as privacy) without specifying who should promote those principles and how, or in which domain they should be promoted and for what reason. A smaller number of documents offered a list of specifications including domain-specific and stakeholder-specific sets of good practices. Among those, a small portion of documents provided explicit recommendations for IRBs or analogous ethics review oversight bodies. In-depth thematic analysis of this subset of documents revealed four major procedural recommendations for IRBs (see appendix 4).

Strengthen oversight function. IRB oversight should be required for big data research even when the research project does not involve the physical (offline) recruitment of human subjects and does not process personally identifiable data or entail direct foreseeable harms to data generators. Furthermore, the IRB's purview should be expanded to monitor the ethical soundness of big data projects throughout the whole data lifecycle. The IRB's control mechanisms should be able to audit each phase of the project, including research planning, data collection, analytics, and results dissemination. The IRBs must inspect if ethical safeguards are in place to protect individual and group-level privacy, autonomy, safety, and the quality and transparency of data management. A few documents argued that IRBs should have the capacity to anticipate or prospectively identify if violations of data access rights might occur and to manage ethical risks associated with data disclosure—particularly when analytic techniques are in use that can allow data processors (or third parties) to reidentify individuals and reveal sensitive information. Given the expertise and independent role of IRBs, they were generally claimed to be well-suited to guarantee an impartial and objective oversight of big data studies.

Improve the review process. IRBs should improve their review process to account for the novel ethical challenges of big data studies. For example, documents noted that IRBs might need to reconsider how to review informed consent procedures in large-scale data-driven projects where traditional informed consent models might be unfeasible (especially in the case of secondary or tertiary data uses).[33] Similarly, documents highlighted that balancing risks and benefits is increasingly complicated in the age of big data, as indirect and informational risks are harder to detect compared to the conventional physical risks of clinical research.[34] Furthermore, risks such as personal data leakages across multiple data cycles are difficult to anticipate prior to data collection and during the ethics review phase. Documents highlighted the need "to create or expand accountable data ethics review processes"[35] and/or develop a new ethical framework specific to big data research. Some documents also suggested the creation of new independent advisory boards whose function should be complementary to that of IRBs.[36]

Build capacity and expand competencies. Documents proposed that IRB members should receive additional training in data science and expand their knowledge of the ethical challenges of big data research. Wherever necessary, IRBs should also consider diversifying their membership to include data scientists and data ethicists. Several documents noted that IRBs are often composed of stakeholders (such as lawyers, physicians, nurses, and laypeople) who rarely have received formal training in computer or data science. Building capacity and expanding competencies are critical to anticipate and promptly identify ethical challenges. Some documents hypothesized that doing so will thereby improve the credibility of IRBs from the point of view of researchers and will increase researchers' willingness to undergo ethics review.[37]

Engage with researchers. IRBs should engage more with researchers and involve them in the ethical evaluation of big data projects. The analyzed documents reported that to achieve this goal, IRBs should have an open dialogue with researchers, sensitize them about the importance of ethically aligned research, and devel-

op facilitated channels for the ethics review of big data projects. At the same time, IRBs should get involved, together with academic ethicists, in the research ethics training of young data scientists. IRBs and researchers should establish a tight collaboration to identify, preempt, and manage ethical risks emerging in health-related big data research. Overall, we recognized high heterogeneity in the way recommendations were carried out by different issuers (see appendix 4).

We note here several study limitations. In the phase of literature retrieval, three typical limitations of gray literature reviews applied: selection bias, the volatile structure of web content, and the documents' heterogeneity. As our search string was written in English and our inclusion criteria included only articles written in any of the languages known by one or more of the researchers, articles written in any other language have not been included. While this limitation is inherent to any literature review, we believe we have minimized it by including articles written in five languages (English, French, German, Greek, and Italian). Another possible source of selection bias is that the Google search engine results are usually returned customized to a specific user and ranked following the number of hits a website received. To anticipate this problem, the search was performed independently by two researchers using separate terminals and IP addresses and in nonpersonalized mode. Google pages were screened until data saturation was reached and acknowledged by both researchers. To minimize subjective bias, the two researchers compared their results and, in cases of divergence, discussed them with motivations until agreement was reached.

The absence of an exhaustive repository of soft-law and policy documents, together with the volatility of their web-based content, might have also affected the review. On the one hand, relevant documents might have gone undetected due to the sensitivity of our search. On the other hand, retrieved documents might, in the future, be removed from the Internet. To minimize the first risk, two coauthors independently screened multiple Google results pages and then crosschecked their results. To address the latter, we retained the original documents in PDF format and created a private repository, which will be shared upon request.

Since we included in the analysis documents that were very diverse in format, content, and quality, this heterogeneity might have affected our thematic analysis. While we are aware of this limitation, we believe that more selective inclusion criteria would have defeated the exploratory purpose of our review. Finally, inductive thematic analyses are also vulnerable to the problem of subjective interpretability by different researchers. This subjective bias is due to the methodological freedom in constructing themes by grouping codes inductively derived from the texts. Although other researchers might have chosen different classification systems, we assessed our thematic classifications iteratively and adapted them along the way to verify their consistency and adherence to the data.

DISCUSSION

The literature review we conducted illustrates a growing corpus of soft-law documents on the ethics review of health-related big data science. The overall number of documents published on this topic increased linearly from year 2012 onwards, indicating a growing interest among regulators and other stakeholders. At the same time, the heterogeneous corpus of documents is indicative of a fragmented ethical and regulatory landscape rather than of an internationally shared framework for ethically aligned big data research. The spectrum of actors involved in this domain is diverse, as it includes, among others, regional (such as the Information and Privacy Commissioner of Ontario), governmental (such as the U. S. Office for Human Research Protections), intergovernmental (such as UNESCO), and supranational (such as the Council of Europe and the European Commission) institutions as well as private companies and NGOs. Very few documents were issued by academic research institutions, despite their direct involvement in research. In contrast, a considerable number of documents were issued by professional associations such as the United Kingdom's Royal Statistical Society and the Internet Association of Privacy Professionals. An even smaller portion of the corpus is represented by independent ethics bodies such as the Nuffield Council on Bioethics and the Italian National Bioethics Committee. However, their documents appeared ethically richer and more detailed compared to the average—an observation that is corroborated by the higher-than-average number of codes identified among these documents.

While it cannot be ruled out that private actors' involvement in big data ethics is indicative of a genuine ethical interest, it has been observed that their proactive guidance efforts have scarce democratic accountability and might raise a risk of undue influence on policymaking, especially when applied to pervasive systems such as data analytics and artificial intelligence.[38] In fact, many large health-related datasets are exclusive property of companies, whose data handling and operational strategies are often hidden by nondisclosure agreements. Given the critical role of private corporations in the data economy, this industry mobilization is necessary to shape an enforceable ethical framework for big data research. At the same time, it raises the quandary of social accountability and the risk that nonstate actors might acquire a quasilegislative power. These problems have particular significance when procedural or substantive conflicts arise between the recommendations provided by, respectively, industry actors and governmental bodies.

Content analysis reveals that most documents provide general normative recommendations about the ethical, legal, and social implications of big data without specifying to which domain these recommendations apply. Moreover, from these documents, it is not clear which actors or bodies should be entitled to promote or enforce these recommendations. Only a minority of documents (33%) specifically addressed IRBs or other ethics review bodies by developing ad hoc recommendations for the review of big data projects. The reason for that is possibly twofold. First, issuer groups such as professional associations, NGOs, and private companies rarely engage with IRBs. Second, big data studies that do not involve human subjects are often perceived as falling outside the purview of ethics review.[39] This interpretation is corroborated by documents such as the Menlo Report[40] and the Data & Society Report,[41] which reveal that researchers involved in data-intensive research typically avoid formal ethics review, as they do not perceive it to be "human subjects research," especially when they rely on secondary deidentified data collections or on corporate-owned databases. This result is consistent with previous studies[42] showing that researchers using big data methods are more likely to bypass IRB review and to adopt self-assessment. While self-regulation approaches are well-suited to ensure scientific freedom, bypassing ethics review via self-assessment is ethically problematic. As the history of biomedical ethics has repeatedly shown, the avoidance of independent ethics review can lead to individual or societal harm and diminish the public's trust in science.[43] This is particularly true for novel areas of science whose ethical boundaries and long-term consequences are still subject to predictive uncertainty.

Both public and private actors focused their recommendations on defining the conditions for ethically sound acquisition, processing, and storage of data. The remarkable frequency of codes related to privacy and informed consent indicates a prominent ethical and practical concern around these themes. Nevertheless, ethics of big data should not be reduced solely to a privacy issue.[44] Previous research has observed that, although privacy is a fundamental topic in big data research, it has been overemphasized to the detriment of other issues.[45] Our findings seem to confirm this observation. Our results also indicate that ethical issues of fairness and data ownership are rarely addressed in current guidance documents. This is concerning given the largely reported risks of bias, discrimination, and informational disenfranchisement associated with algorithms and big data analytics.[46] Our results are consistent with previous studies about the governance of artificial intelligence (AI) technologies that found interpretative differences and a lack of actionable requirements for the promotion of fairness and justice in the use of these technologies.[47] These results indicate that attention is missing in this still-developing area of research ethics, and they attest to persistent uncertainty on how fairness and justice considerations should be addressed in the age of big data and AI. We argue that more detailed normative guidance is needed in this regard.

The high level of interconnectedness among different ethical macrothemes highlights that ethical issues are not in silos but are intimately intertwined. This makes the ethics review of big data projects a complex and multifaceted process that involves not only scrutinizing ethical codes and methods but also inspecting technical requirements, addressing epistemological considerations, and anticipating societal implications. Results suggest that IRBs should exercise their role of essential control systems evaluating and balancing the

different faces of each issue, which might require expanded purview and diversified expertise.

Given the fragmentation and heterogeneity of the current landscape of guidance documents, it is unlikely to reduce uncertainty among researchers and IRBs regarding the ethics review of health-related big data studies. Nonetheless, our results revealed a recurrence of four major procedural recommendations for IRBs. These recommendations address how IRBs should improve their review activities, strengthen their competencies, and revise some of their established practices.

First, documents identified a need for more comprehensive oversight strategies, especially by expanding the purview of IRBs to require formal ethical assessment of data-intensive studies even when they do not involve the recruitment of human subjects or operate on publicly available data repositories. At the same time, researchers should interpret ethics review not as a waiver of their responsibility but, rather, as an essential quality control of their research. While expanding the purview of IRBs might require new legislation, encouraging researchers to undergo an ethics review on a voluntary basis might be a temporary measure to improve ethical safeguards. Voluntary submissions for review can be incentivized through awareness-sensitive campaigns about the ethical implications of big data among researchers and by fast-track review procedures for projects that ensure certain technical requirements.

Second, documents urged IRBs, research institutions, and science regulators to improve the ethics review process and formalize a coherent ethical review framework for the evaluation of big data projects. Documents observed that research ethics paradigms developed for offline research are hardly transferable to data-driven research in absence of calibration. For example, research aimed at mining health-related data might have challenging implications for conceptual milestones of human research ethics such as the notion of minimal risk. Unlike conventional research involving human subjects, big data research involving human-related data might not pose direct risks for the physical integrity of research participants. However, the last few years have borne out the fact that poorly managed datasets can have harmful consequences for human subjects in terms of mental well-being, harm to reputation, unfair treatment, and discrimination or other forms of informational risk and dignitary harm.[48] Consequently, the standards of minimal risks developed for clinical research are hardly applicable to the big data domain if significant conceptual and normative adjustments are not performed.

Third, documents highlight the importance of empowering IRBs with the relevant expertise to account for the computational and ethical complexity of big data studies. IRB members trained in medicine, psychology, law, or traditional research ethics might lack the relevant expertise to determine whether, for instance, a certain project is deploying safeguards to avoid algorithmic discrimination, if the machine learning models used for decision-making are amenable to ex ante and post hoc inspection, or if group-level privacy risks can arise from the combination of differently structured data sources. Documents suggest that this epistemological gap can be filled with a two-pronged approach: by diversifying the IRB's composition and through capacity-building strategies. To diversify their composition, IRBs should consider appointing individuals with expertise in computer science, data analytics, statistics, and data ethics. Furthermore, they should consider the organization of training programs or other educational and capacity-building activities.

Expanding the IRB purview and their members' expertise is a requirement grounded on the assumption that IRBs should be the relevant oversight body of big data research. This assumption was not shared unanimously. A few documents addressed the issue of whether IRBs should be the oversight body accountable for big data research at all.[49] For example, data protection officers were proposed as complementary oversight resources. The creation of novel oversight bodies such as data boards was also proposed as an adaptive governance solution to the big data ethics conundrum.

Finally, documents highlighted the importance of sensitizing researchers and other relevant actors (for instance, technology developers, data analysts, advertisers, insurers, and physicians) about data ethics. The persistent absence of an agreed-upon ethical framework for big data research might perpetuate uncertainty between both researchers and IRBs and could result in divergent approval decisions. Raising awareness within the research community can help reduce this uncertainty through proactive measures such as the development of

codes of ethics and professional conduct (as done by the Association for Computing Machinery and the British Computer Science Association), research roadmaps (as done by the Association of the British Pharmaceutical Industry), or best practices (as done by the Health IT Policy Committee of the United States). Any development of an ethical framework for big data research, however, cannot disregard the active involvement of IRBs in decision-making. On the contrary, research on the views, needs, and attitudes of IRB members is highly necessary to set an evidence-based, empirically informed agenda for big data and research ethics.

Despite the prevalence of the above-listed recurrent themes across documents, there is still much uncertainty about how the recommendations should be implemented. For instance, it is not clear yet how, in practice, IRBs should improve the ethics review process and which recommendation should be implemented first. Whether IRBs should be the bodies devoted to assessing big data projects at all is still debatable. Alternatives might involve universities providing ethical requirements to their researchers who collect, store, or use big data. Additionally, peer-reviewed journals might set the rule to reject all those publications that do not follow specific ethical procedures and criteria. Another option could involve producing new legislation that includes new research ethics best practices. When advancing this ethical discussion, it is critical that IRBs are not considered passive recipients of guidelines but are actively involved in the norm-development process. To achieve this aim, qualitative studies assessing the views, needs, and attitudes of IRBs as well as collaborative approaches to guideline development are highly needed. ♦

SUPPORTING INFORMATION

The figures and appendices are available in the "Supporting Information" section for the online version of this article and via *Ethics & Human Research*'s "Supporting Information" page: https://www.thehastingscenter.org/supporting-information-ehr/.

Agata Ferretti, MA, MSc, *is a PhD candidate in Bioethics at the Health Ethics and Policy Lab, Swiss Federal Institute of Technology in Zurich*; **Marcello Ienca, PhD, MA, MSc,** *is a senior researcher at the Health Ethics and Policy Lab, Swiss Federal Institute of Technology in Zurich;* **Samia Hurst, PhD,** *is a professor of medical ethics and director of the Institute for Ethics, History, and the Humanities & Department of Community Health and Medicine, University of Geneva, Switzerland*; *and* **Effy Vayena, PhD,** *is a professor of bioethics and director of the Health Ethics and Policy Lab, Swiss Federal Institute of Technology in Zurich.*

REFERENCES

1. Hershey, N., "IRB Jurisdiction and Limits on IRB Actions," *IRB: Ethics & Human Research* 7, no. 2 (1985): 7-9; Cook, A. F., and H. Hoas, "Protecting Research Subjects: IRBs in a Changing Research Landscape," *IRB: Ethics & Human Research* 33, no. 2 (2011): 14-19.

2. Ta, V. D., C. M. Liu, and G. W. Nkabinde, "Big Data Stream Computing in Healthcare Real-Time Analytics," in *2016 IEEE International Conference on Cloud Computing and Big Data Analysis* (Chengdu: IEEE, 2016): 37-42.

3. Marx, V., "The Big Challenges of Big Data," *Nature* 498, no. 7453 (2013): 255-60; Greely, H. T., K. M. Ramos, and C. Grady, "Neuroethics in the Age of Brain Projects," *Neuron* 92, no. 3 (2016): 637-41; Landhuis, E., "Neuroscience: Big Brain, Big Data," *Nature* 541 (2017): 559-61; Insel, T. R., "Digital Phenotyping: Technology for a New Science of Behavior," *Journal of the American Medical Association* 318 (2017): 1215-16.

4. Vayena, E., and U. Gasser, "Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine," in *The Ethics of Biomedical Big Data* (Cham, Switzerland: Springer, 2016): 17-39.

5. Ienca, M., E. Vayena, and A. Blasimme, "Big Data and Dementia: Charting the Route Ahead for Research, Ethics, and Policy," *Frontiers in Medicine* 5, no. 13 (2018): doi.org/10.3389/fmed.2018.00013.

6. Mazzocchi, F., "Could Big Data Be the End of Theory in Science? A Few Remarks on the Epistemology of Data-Driven Science," *EMBO Reports* 16, no. 10 (2015): 1250-55.

7. Metcalf, J., and K. Crawford, "Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide," *Big Data & Society* 3, no. 1 (2016): doi:0.1177/2053951716650211.

8. Vayena, E., et al., "Elements of a New Ethical Framework for Big Data Research," *Washington and Lee Law Review Online* (2016): doi.org/10.3389/fmed.2018.00013.

9. European Commision, *Ethics for Researchers: Facilitating Research Excellence in FP7* (Brussels: European Commission, 2013), http://ec.europa.eu/research/participants/data/ref/fp7/89888/ethics-for-researchers_en.pdf.

10. SAMS, "Research with Human Subjects," Swiss Academy of Medical Sciences, (2015), https://www.sams.ch/dam/jcr:33181182-5ed6-4004-addc-86535089cfd9/handbook_sams_research_2015.pdf.

11. Holm, S., and T. Ploug, "Big Data and Health Research—the Governance Challenges in a Mixed Data Economy," *Journal of Bioethical Inquiry* 14, no. 4 (2017): 515-25.

12. Ienca, M., et al., "Considerations for Ethics Review of Big Data Health Research: A Scoping Review," *PloS One* 13, no. 10 (2018): doi.org/10.1371/journal.pone.0204937; Vayena, E., and A. Blasimme, "Health Research with Big Data: Time for Systemic Oversight," *Journal of Law, Medicine & Ethics* 46, no. 1 (2018): 119-29; Christen, M., et al., "On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project," in *The Ethics of Biomedical Big Data,* ed. Mittelstadt, B. D., L. Floridi (Cham, Switzerland: Springer, 2016): 199-218; Ioannidis, J. P., "Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research," *American Journal of Bioethics* 13, no. 4 (2013): 40-42; Knoppers, B. M., and A. M. Thorogood, "Ethics and Big Data in Health," *Current Opinion in Systems Biology* 4 (2017): 53-57.

13. El Emam, K., S. Rodgers, and B. Malin, "Anonymising and Sharing Individual Patient Data," *BMJ* 350 (2015): h1139.

14. Wang, Y., and M. Kosinski, "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images," *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246.

15. "Cambridge Analytica Controversy Must Spur Researchers to Update Data Ethics," editorial, *Nature* 555 (2018): doi:10.1038/d41586-018-03856-4.

16. EurWORK (European Observatory of Working Life), "Soft Law," Eurofound, May 4, 2011, https://www.eurofound.europa.eu/observatories/eurwork/industrial-relations-dictionary/soft-law; Sossin, L., and C. W. Smith, "Hard Choices and Soft Law: Ethical Codes, Policy Guidelines and the Role of the Courts in Regulating Government," *Alberta Law Review* 40, no. 4 (2003): 867-93.

17. Ienca et al., "Considerations for Ethics Review."

18. Dereli, T., et al., "Big Data and Ethics Review for Health Systems Research in LMICs: Understanding Risk, Uncertainty and Ignorance—and Catching the Black Swans?," *American Journal of Bioethics* 14, no. 2 (2014): 48-50.

19. McGrath, Y., et al., *Review of Grey Literature on Drug Prevention among Young People* (London: NICE, 2006); Favin, M., et al., "Why Children Are Not Vaccinated: A Review of the Grey Literature," *International Health* 4, no. 4 (2012): 229-38.

20. Higgins, J. P., and Green, S., "Cochrane Handbook for Systematic Reviews of Interventions," (2008), https://training.cochrane.org/handbook.

21. Alberani, V., P. D. C. Pietrangeli, and Mazza, A., "The Use of Grey Literature in Health Sciences: A Preliminary Survey," *Bulletin of the Medical Library Association* 78, no. 4 (1990): 358.

22. Aina, L., "Grey Literature and Library and Information Studies: A Global Perspective," *International Journal on Grey Literature* 1, no. 4 (2000): 179-82.

23. McGrath et al., *Review of Grey Literature.*

24. Daniulaityte, R., et al., "Qualitative Epidemiologic Methods Can Improve Local Prevention Programming among Adolescents," *Journal of Alcohol and Drug Education* 48 (2004): 73-84.

25. Benzies, K. M., et al., "State-of-the-Evidence Reviews: Advantages and Challenges of Including Grey Literature," *Worldviews on Evidence-Based Nursing* 3, no. 2 (2006): 55-61.

26. McGrath et al., *Review of Grey Literature*; Lawrence, A., "Influence Seekers: The Production of Grey Literature for Policy and Practice," *Information Services & Use* 37, no. 4 (2017): 389-403; Simkhada, P., et al., "Chasing the Grey Evidence: A Standardised Systematic Critical Literature Review Approach," paper presented at the GL-conference series, 2005, http://www.opengrey.eu/item/display/10068/697851.

27. EurWORK, "Soft Law"; Sossin and Smith, "Hard Choices and Soft Law."

28. Herzog, C., C. Handke, and E. Hitters, "Analyzing Talk and Text II: Thematic Analysis," in *The Palgrave Handbook of Methods for Media Policy Research*, ed. H. Van den Bulck et al. (Basingstoke, England: Palgrave Macmillan, 2019), 385-401.

29. Braun, V., and V. Clarke, "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology* 3, no. 2 (2006): 77-101.

30. Kramer, A. D., J. E. Guillory, and J. T. Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks," *Proceedings of the National Academy of Sciences* (2014): 8788-90.

31. Jagadish, H. V., et al., "Big Data and Its Technical Challenges," *Communications of the ACM* 57, no. 7 (2014): 86-94.

32. "ISO/IEC 27001:2013," ISO (International Organization for Standardization), https://www.iso.org/standard/54534.html.

33. U.K. Data Service, *Big Data and Data Sharing: Ethical Issues* (updated February 2017), https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf; Information Commissioner's Office, *Big Data, Artificial Intelligence, Machine Learning and Data Protection* (2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

34. Council of Europe, *Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in a World of Big Data* (Strasbourg, France: Council of Europe, 2017), https://rm.coe.int/16806ebe7a; "Attachment A: Human Subjects Research Implications of 'Big Data' Studies," Office for Human Research Protections, 2015, https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2015-april-24-attachment-a/index.html.

35. Tene, O., and J. Polonetsky, "Big Data for All: Privacy and User Control in the Age of Analytics," Northwestern Journal of Technology and Intellectual Property 239 (2013): article 1.

36. Council of Europe, *Guidelines on the Protection of Individuals*; International Association of Privacy Professionals,

Building Ethics into Privacy Frameworks for Big Data and AI (Strasbourg, France: Council of Europe, 2017), https://rm.coe.int/16806ebe7a.

37. Boyd, D., E. F. Keller, and B. Tijerina, "Supporting Ethical Data Research: An Exploratory Study of Emerging Issues in Big Data and Technical Research," working paper, Data & Society, August 4, 2016, https://www.datasociety.net/pubs/sedr/SupportingEthicsDataResearch_Sept2016.pdf; Nuffield Council of Bioethics, "Response to the Science and Technology Select Committee (Commons) Inquiry: The Big Data Dilemma," September 2015, http://nuffieldbioethics.org/wp-content/uploads/Big-Data-dilemma-Nuffield-Council-on-Bioethics-September-2015.pdf.

38. Benkler, Y., "Don't Let Industry Write the Rules for AI," *Nature* 569 (2019): 161.

39. "Cambridge Analytica Controversy Must Spur Researchers to Update Data Ethics," editorial, *Nature* (March 27, 2018), https://www.nature.com/articles/d41586-018-03856-4.

40. The Menlo Report, 2012, https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf.

41. Supporting Ethical Data Research: An Exploratory Study of Emerging Issues in Big Data and Technical Research, 2016, https://www.datasociety.net/pubs/sedr/SupportingEthicsDataResearch_Sept2016.pdf.

42. Samuel, G. N., and B. Farsides, "Public Trust and 'Ethics Review' as a Commodity: The Case of Genomics England Limited and the UK's 100,000 Genomes Project," *Medicine, Health Care and Philosophy* (2017): 159-68; Metcalf and Crawford, "Where Are Human Subjects in Big Data Research?"

43. Gracia, D., "History of Medical Ethics," in *Bioethics in a European Perspective,* ed. H. T. Have and B. Gordjin (Cham, Switzerland: Springer, 2001), 17-50.

44. Mason, P. H., *The Ethics of Biomedical Big Data,* ed. B. D. Mittelstadt and L. Floridi (Cham, Switzerland: Springer International Publishing, 2016); Ienca et al., "Considerations for Ethics Review."

45. Mittelstadt, B. D., and L. Floridi, "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts," *Science and Engineering Ethics* 22, no. 2 (2016): 303-41.

46. Cuquet, M., and A. Fensel, "The Societal Impact of Big Data: A Research Roadmap for Europe," *Technology in Society* 54 (2018): 74-86; Knoppers and Thorogood, "Ethics and Big Data in Health."

47. Jobin, A., M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389-99.

48. "Cambridge Analytica Controversy Must Spur Researchers to Update Data Ethics," editorial, *Nature*.

49. Council of Europe, *Guidelines on the Protection of Individuals*; International Association of Privacy Professionals, *Building Ethics into Privacy Frameworks*.