Scraping the Web for Public Health Gains: Ethical Considerations from a 'Big Data' Research Project on HIV and Incarceration

Stuart Rennie*, Mara Buchbinder, and Eric Juengst, UNC Bioethics Center, Department of Social Medicine, University of North Carolina at Chapel Hill

Lauren Brinkley-Rubinstein, Center for Health Equity, Department of Social Medicine, University of North Carolina at Chapel Hill

Colleen Blue and David L. Rosen, Institute for Global Health and Infectious Diseases, University of North Carolina at Chapel Hill

*Corresponding author: Stuart Rennie, Department of Social Medicine, Center for Bioethics, University of North Carolina at Chapel Hill, 333 MacNider Hall, 333 S Columbia Street, Chapel Hill, NC 27519, USA. Tel.: 919-962-7594; Email: stuart_rennie@med.unc.edu

Web scraping involves using computer programs for automated extraction and organization of data from the Web for the purpose of further data analysis and use. It is frequently used by commercial companies, but also has become a valuable tool in epidemiological research and public health planning. In this paper, we explore ethical issues in a project that "scrapes" public websites of U.S. county jails as part of an effort to develop a comprehensive database (including individual-level jail incarcerations, court records and confidential HIV records) to enhance HIV surveillance and improve continuity of care for incarcerated populations. We argue that the well-known framework of Emanuel et al. (2000) provides only partial ethical guidance for the activities we describe, which lie at a complex intersection of public health research and public health practice. We suggest some ethical considerations from the ethics of public health practice to help fill gaps in this relatively unexplored area.

Introduction

The World Wide Web can be regarded as the largest database ever created in human history. Precise estimates of its magnitude—in terms of storage, communication and computation—are a matter of debate (Pappas, 2016), but it is safe to say that the volume, speed and variety of data give the internet unprecedented potential to advance important social goals. As human actions are increasingly captured by digital technology, online data are becoming a highly valued information source for researchers of all stripes. In medicine and public health, many hope that vastly increased volumes of patient data will help overcome existing knowledge gaps, lead to health innovations and improve health outcomes.

One pervasive method of gathering digital data is the practice of 'web scraping'. The massive amount of data available on the Web means that effective data collection and processing cannot be manually conducted by individual researchers or even large research teams. Web scraping, an alternative to manual data collection, entails the use of computer programs for automated extraction and organization of data from the Web for the purpose of further data analysis and use (Krotov and Silva, 2018). Commercial companies are heavily reliant on web scraping to collect, for example, data on consumer preferences (e.g., in product reviews) and business competitors (e.g., prices) in real time to inform goals and strategy.

Web scraping has also become a valuable tool in epidemiological research and public health planning. With massive, publicly accessible health-related data available on the internet, epidemiologists increasingly need to be trained in computer programming, including web scraping (Mooney *et al.*, 2015). Web scraping for public health purposes is not limited to health data; it can also include data of potential biomedical or public health significance, such as social media posts about meals or

doi:10.1093/phe/phaa006
Advance Access publication on 11 March 2020

eating habits, records of court actions, or traffic patterns (Vayena and Gasser, 2016; Richterich, 2018).

However, as its evocative name suggests, web scraping is not necessarily a benign procedure. In fall 2017, the San Francisco-based company Strava, described as a 'social network for athletes', announced an update to its global heat map of user activity that visually tracks the movements of users wearing Fitbits or other fitness trackers. Journalists and activists have raised ethical concerns regarding the fact that data scraped from Strava's website could be used to track the location of military or intelligence personnel and identify individual users when combined with other information. These events are prompting the US military to rethink its policies regarding use of fitness trackers, which were previously encouraged to promote physical activity (Hsu, 2018).

Considering such potential misuses, it is not always clear how to interpret and apply privacy standards and laws that were established prior to our highly interconnected, digitalized world (Gold and Latonero, 2018). What level of control should individuals have over information posted by others about them? Are some individuals more at risk than others from research that 'scrapes the web' to glean publicly accessible information about them? What safeguards should be in place to prevent social harms that might be associated with the use of individualized online data in research and public health contexts? What other considerations should inform the just use of this enormous resource for research purposes?

To explore these questions in the context of a specific illustrative case, we discuss the ethics of an ongoing study in which we are 'scraping' public websites of US county jails to create a database of individual-level jail incarcerations. In collaboration with the North Carolina Health Department, we will combine these jail incarceration records with existing confidential HIV records to create a database that could, as we explain below, (i) inform new, enhanced forms of HIV surveillance incorporating jail populations and (ii) potentially contribute to future public health approaches for improving care for incarcerated populations. As such, this case study raises questions about the use of web scraping in both research and public health contexts. In the next section, we describe the case in some detail, and apply the widely used Emanuel *et al.* (2000) research ethics framework to raise and examine ethical concerns related to use of web scraping in public health research. The Emanuel *et al.* framework was initially developed as guidance for the design and conduct of *clinical* research, but is often extended to other types of biomedical research such as public health research. We use it as our point of departure in order to ask how well it can help address raised by *translational* public health research that sits at the intersection of public health research and public health practice, like the web scraping activities of our case study. We argue that certain ethically salient aspects of 'big data' research in public health contexts are neglected by this influential framework. This is in part because of the ways in which such research would inform public health surveillance practices, which would also depend on ongoing web scraping to succeed. We conclude by suggesting how considerations from the ethics of public health practice can help to address these intersectional blind spots.

Case Study: Leveraging Big Data to Understand and Improve Continuity of Care among HIV-Positive Jail Inmates

HIV medications have been definitively shown to improve patients' health and to prevent onward transmission of the virus (Cohen *et al.*, 2011). And yet, in the USA, 40 per cent of people diagnosed with HIV are not retained in care, and therefore, do not have access to HIV medications (Centers for Disease Control and Prevention, 2018). In this context, improving access to HIV care, including access to medications, is a central challenge in combatting the US HIV epidemic.

In 2017, our team began a National Institutes of Health-supported research project to develop a data system to improve HIV care for persons living with HIV in North Carolina who have had periods of incarceration in jail. Based on 2006 data, it was roughly estimated that among all adults in the USA infected with HIV, about one in six annually spend time incarcerated, mostly in local jails (Spaulding *et al.*, 2009). Previous studies have demonstrated the disruptive impact of prison incarceration on HIV treatment and care (Iroh *et al.*, 2015), but currently very little is known about access to medical services among HIV-positive persons before, during or after jail incarceration.

The major aim of our study is to improve continuity of care for justice-involved persons with HIV by improving estimates of the number of people with HIV who are passing through jails in North Carolina and by better understanding their engagement in care before, during and after incarceration. However, several challenges confronted us as we considered different study designs. Given the large number of county jails in the state (97) and the difficulties of recruiting jails as study sites, we determined that a prospective design across multiple sites was not practical. We were aware that some of the

data necessary to answer our question did exist—the state Division of Public Health conducts HIV disease surveillance in which it uses a number of different data sources to determine if people living with HIV are routinely engaged in care. At the same time, there were no existing data sources to address continuity of care for people living with HIV in jails. The state's public health data are integrated into a single database, and could in principle be linked to a database of jail incarceration records, stripped of identifiers and analyzed anonymously for our purposes. The barrier to this approach, however, is that jails operate independently, at the county level, and no single database of jail incarcerations is available to researchers.

Nevertheless, 29 of 97 jails in the state have public websites that provide some information about who is currently incarcerated in their facilities. These jails account for about half of all jail inmates in the state, or about 200,000 people each year. Although the jail websites can vary in their content and layout, they generally include incarcerated persons' names, age or date of birth, arrest charge and date when the incarceration began. To create a database of jail incarcerations, we are web scraping these sites daily. The resulting individual incarceration records will be linked by county, name and date of birth (if available) or age, to a restricted set of state court records, which contain a more robust set of identifiers for people who have pending charges, including date of birth (if missing from the jail data) and partial social security number. As our research project unfolds, this enhanced set of identifiers will then be used by our partners within the state Division of Public Health to link the jail-court records to the state's confidential HIV database, to create a deidentified statewide database of incarcerations involving HIV-infected individuals.

The resulting database can be used to further the goals of public health surveillance in a number of ways, even with all personal identifiers removed. These include providing more precise information about: the burden of HIV-positive inmates in each jail sampled, the length of incarceration for HIV-positive inmates and the patterns of how persons living with HIV access HIV-related care before and after they are in jail. This enhanced surveillance information could be very useful for monitoring purposes and the allocation of resources for medical care among HIV-positive persons who are in and out of jail in North Carolina.

This database could also inform what is called a Data to Care (D2C) approach. D2C involves the use of HIV surveillance data routinely collected by state and local health departments to identify out-of-care individuals and re-link or re-engage them in care. Using this type of approach, state health departments could be notified in real time when a person living with HIV entered jail. This would enable the health department to engage that person in care while in jail, or to connect the individual to health-care providers upon release. This deployment of surveillance data is a shift from its traditional use (i.e. descriptive and monitoring purposes), and has been partly driven by biomedical advances in HIV treatment and prevention (Sweeney *et al.*, 2013). The United States Centers of Disease Control and Prevention (CDC) recently included D2C as a condition of funding for state surveillance efforts and it is currently being implemented in many states (Centers for Disease Control and Prevention, 2018). In other words, a systematic integration of jail, court and public health surveillance data could provide state public health agencies with a powerful tool to identify HIV-positive persons who have been jailed, and help them access medical care.

In linking public jail rosters to individualized court records for comparison against the state health department's confidential HIV database, both our research project and the potential use of this system by the health department and the jails to enhance continuity of care will raise many questions about how best to protect human rights and ensure fairness in the context of web scraping-based surveillance. What is unique about this research project is that it scrapes personal identifiers from websites to develop a database that could contribute to health interventions with individuals (i.e., re-engaging them in HIV care). Most public health initiatives involving web scraping have a population rather than individual-level focus. For example, HealthMap scrapes data about disease outbreaks in real time worldwide from multiple digital sources (such as news feed aggregators, Twitter and agencies such as the World Health Organization) in a Google Map-inspired format. But no personal identifiers are gathered, and its purpose is to provide surveillance information rather than support interventions (HealthMap, 2006). Alternatively, Facebook has created a system that monitors its users' posts for language deemed to convey an imminent risk of suicide; when such a user is identified, first responders can be prompted to reach out to the individual to offer support (Goggin, 2019). However, this project does not employ web scraping as Facebook is utilizing data posted to its own website.

One of the most influential frameworks for thinking through ethical issues in biomedical research is the one proposed by Emanuel *et al.* (2000). Applying this framework to web scraping research helps to illuminate some important ethical considerations in designing such studies and implementing their results in public health

programs. But just as importantly, that framework, designed as it is for conducting clinical trials, also fails to capture other ethical considerations raised by the implementation of web scraping for public health surveillance, which will also be critical to address in applying research findings to public health practice.

Applying the Emanuel Framework to Web Scraping in Public Health Research

The Emanuel *et al.* framework consists of eight principles widely held to be the default ethical requirements for the development, implementation and review of clinical research protocols involving human subjects. The framework has been adapted for use in developing countries (Emanuel *et al.*, 2004) for the review of social science research (Wassenaar and Mamotte, 2012), as well as in specific areas of research, such as HIV phylogenetic studies (Mutenherwa *et al.*, 2019). The requirements are considered universal by Emanuel *et al.*, in the sense that they express widely recognized and accepted moral norms for research, although their use and interpretation can be shaped by contextual and cultural factors. Below, we briefly explain the requirements and relate them to the use of web scraping in our case study.

Social Value

To be ethically justified, it is necessary for a research study to contribute new information that could potentially lead to improvements in health and well-being. This could come in the form of hypothesis-testing, evaluations of interventions or epidemiological studies to help develop interventions. A research study without social value is ethically unjustified because it wastes valuable resources and exploits research participants by exposing them to risks without prospect of social or scientific benefits. The main social value of the research described in our case study is its potential to help improve care for persons living with HIV who become incarcerated. Disruptions in HIV care, and subsequent failure to achieve viral suppression, is an important medical and public health problem. In this case, research involving web scraping has potential social value because it contributes epidemiological information to a database aiming to improve HIV surveillance and continuity of care.

Scientific Validity

A research study with potential social value can nevertheless be ethically problematic if its methods are not sound and the resulting data are unreliable. According to Emanuel *et al.*, the hallmarks of scientific validity are a clear scientific objective and the use of an accepted methodology (including data analysis plan) appropriate for answering the research questions. Data validity has been a concern among those conducting public health research by combining large datasets. While some argue that using higher volumes of data may help avoid some methodological problems inherent in smaller sample sizes, large size by itself does not resolve other forms of error and bias (Chiolero, 2013).

In our study, web scraping jail websites poses a number of methodological challenges. The rapid turnover of inmates at jails, as well delays by jails in updating their websites, means that it is difficult to have a complete record of everyone who has passed through the jails, even when scraping jail websites every few hours. In addition, some inmates use aliases, which can complicate efforts to link information to one and the same individual. We have partly addressed this issue by including in our linkage process the aliases of all active defendants in the court data.

Decisions also need to be made as to what information should or should not be scraped from jail websites, as some web-published data may be superfluous to the needs of the project. For example, we discussed the possibility of whether 'mugshots'—forensic photographic portraits—could ever be effectively used in differentiating between similar entries on a jail website (e.g. two entries with the same last name but slightly different spellings of a similar first name). Ultimately, we decided that the resources necessary for manual or automated inspection of mugshots were beyond the scope of the project, and therefore, we decided against collecting these images despite their potential use in enhancing our data validity. Moving forward, we will continue to assess scientific validity as we monitor and improve upon our process of linking data between the jails, court system and state public health department.

Fair Subject Selection

Inclusion and exclusion of research participants should not just advance scientific goals, but also be responsive to considerations of fairness. This requirement has several different dimensions, but in the context of health research, it concerns the equitable distribution of the burdens and benefits of research participation. In the past, research involving those in prison or jail was often

exploitative in that it was designed to provide benefits for others, i.e. the general, non-incarcerated population (Gostin, 2007). Selecting incarcerated individuals for research can be fair if the research responds to health problems specifically faced by this population and they are likely to benefit from the results. Our study collected incarceration data for the purpose of improving HIV care services for those who have been jailed. Fair subject selection, therefore, at least partly depends on the burdens and risks related to being in the database rather than not (see 'favorable risk-benefit ratio' below). In our study, inclusion in the database is primarily a result of whether the individual is incarcerated in a jail that chooses to publish a website with an inmate roster. Notably, about two-thirds of jails in our state (accounting for about half of all incarcerated persons) do not publish a website roster, and these tend to be lower resourced jails. Accordingly, people incarcerated in the lower resourced jails may be the least likely to be included in our research project. If future interventions based on this research lead to improved HIV care, those in lower resourced jails may be (at least initially) excluded from them.

In health-related studies involving technology (e.g. mobile phones), participation bias and fair subject selection are persistent challenges when access to the technology is limited. In our case study, it is a matter of institutions (i.e. jails) not publishing data online, rather than individuals not having the requisite devices. This reveals a dimension of fairness that does not enter the Emanuel *et al.* framework: power asymmetries, as they manifest in access to, use, and impact of information and communication technologies. In contemporary societies, data on individuals are routinely collected, stored and used by powerful institutions, both governmental and commercial. The purpose of such data collection and the motives of those who hold and use it are proper objects of ethical inquiry. In our case, inclusion or exclusion of potential research participants may depend not on scientific criteria, but on what information jails are willing to release; important data may be made inaccessible by jails in order to prevent public exposure of problems in the criminal justice system. In this context, collection of jail data (including use of web scraping) may be morally imperative whether jail authorities agree with it or not.

Favorable Risk–Benefit Ratio

Research invariably involves a certain degree of risk, and the ethical justification of exposure to risk depends on the relationship between risks to participants and benefits to the individual and society. Studies should be designed to minimize potential risks to participants without compromising the scientific validity of the research, while enhancing (when possible) potential benefits for individuals and society. The question of risk in our study is complicated. On the one hand, potential benefits for incarcerated individuals and society is high, particularly if the research contributes to more effective surveillance and increases re-engagement of the affected population into care. One could also argue this study aims to minimize risk, given that those who currently live with HIV but are not virally suppressed pose health risks to themselves and others. In addition, gathering information on jail websites that is already in the public domain would traditionally be understood as minimal risk, because the potential infringement of privacy piggybacks on an existing convention (at least in the USA).

On the other hand, whether the benefit–risk relationship is favorable depends heavily on the efficacy of the safeguards in place to protect sensitive information from inappropriate disclosure. It also depends on context, such as whether one is considering the use of data in enhanced surveillance or future use in a D2C approach. In the use of web scraping for enhanced surveillance, personal identifiers are being collected, even though the data will be anonymized for surveillance purposes. Furthermore, the information collected is sensitive and being processed in ways that could increase stigma: persons in the dataset have their HIV infection linked to formal criminal investigation and are characterized as appropriate targets of public health action. In addition, the meaning of 'publicly available' should be regarded critically. While it is true that ordinary citizens can gather some information about the incarcerated by visiting jail websites, web scraping can gather information in greater magnitude that can be used to generate insights about the incarcerated beyond the capacities of lay users (Nissenbaum, 2009). Depending on how the data are used, web scraping can increase the visibility of a person's incarceration history. In the enhanced surveillance use of the jail website data, where personally identifying information is removed, there is less risk of social harms to individuals while gaining potential individual and public health benefits. However, in the D2C use of the web scraped data, the risk of social harm to individuals could increase as their incarceration status is explicitly linked to court data and HIV status, and identifiers are retained to facilitate engagement in care.

Independent Review

Ethical review of research by third parties uninvolved in the research is meant to act as a counter to potential

biases of researchers and to provide public assurance that the rights and welfare of research participants are adequately protected. The review should include assessment of scientific validity, social value, risks and benefits, the informed consent process and community engagement. None of this is unique to research involving web scraping. However, while web scraping has been practiced for decades, it is a relatively new approach to gathering information in health research and, as can happen with unfamiliar methods, members of ethics committees may reject studies that involve web scraping out of an abundance of caution. This is particularly likely in countries, such as the USA, where incarcerated populations are given extra research protections. Conversely, research ethics committees that are more familiar with the routine use of web scraping in other contexts may prematurely approve such studies without insight into potential risks and risk mitigation strategies. A challenge for research ethics committees with respect to web scraping and its applications is that protections for data have become highly technical; gone are the days when locking hard copies of data in cabinets or even basic file encryption is sufficient. Ethical evaluation of data safeguards increasingly requires input from information technology experts.

It is noteworthy that US Federal regulations have become less strict for studies not directly engaging human participants at the same time that researchers are increasingly using 'big data' resources and when many in the research ethics community are concerned about the lack of protection offered by informed consent processes and the challenges of preserving anonymity (Barocas and Nissenbaum, 2014). This means that if there are ethical issues in studies using web scraping that are not captured by current regulations, much depends on research ethics committee expertise in sophisticated data processes and protections. While Emanuel *et al.* rightly assert that research ethics committee members should be 'competent', the competence required to assess ethical protections in web scraping is different than merely understanding research methodologies or ethics.

Informed Consent

The requirement for informed consent is based on the principle of respect for persons, where 'respect' is understood in terms of individuals having control over the decision to be part of a research study. The challenges of gaining valid informed consent in research generally, where individuals have adequate understanding of what participation involves and agree without inappropriate influence, are well-known (Grady, 2015). In our study, individuals whose jail data are being gathered through web scraping did not consent to giving their information to jail authorities, having this information collected from jail websites for research purposes, or potential future uses. The same is true for their court and public health data. For its part, our Institutional Review Board considered our study minimal risk, approved it by expedited review in collaboration with a prison representative, and authorized a waiver of the requirement of informed consent on the basis of US federal regulations (45 CFR 46.116[d]).

From an ethical standpoint, one could argue against obtaining informed consent from incarcerated individuals in a number of ways. As mentioned earlier, the jail website data were already publicly available, and the benefits of the study appear to outweigh the risks. Furthermore, the purpose of the research was to contribute to surveillance efforts, and health surveillance is commonly conducted without individual consent when doing so clearly promotes public welfare. In addition, obtaining individual consent could undermine the scientific validity and social value of surveillance by introducing participation bias. Finally, obtaining individual consent from approximately 200,000 persons within the time span of our study would be practically impossible.

While some may find these reasons persuasive, there are still some loose ends. The potential D2C applications of our study, which would require identifiable information, raise questions about the appropriateness of waiving the requirement of informed consent. Imagine that the fully identified version of our database was to be used by public health agents to approach formerly incarcerated individuals who appear to have discontinued HIV care services. In that imagined scenario, the formerly incarcerated individuals may wonder how they came to be identified. The answer would be that they were contacted on the basis of an enhanced surveillance database. Will the beneficiaries of such enhanced surveillance and outreach feel that their privacy has been violated, considering how their data—some of which is sensitive—has been collected from various sources and used by agencies without their permission? Or will they appreciate the assistance? Sweeney *et al.* (2013) provide some evidence that D2C results in increased uptake of HIV services and (for the most part) acceptance of being contacted. While these data did not focus on incarcerated individuals, and does not put the ethical issue of nonconsent to rest, it at least suggests that health-related web scraping/surveillance/outreach initiatives may not necessarily erode the public's trust.

Consent of web scrapers and web hosts must also be considered. While some web scrapers contact website hosts to request specific datasets, others simply scrape websites for the data they desire without asking permission. To some extent, website hosts can determine the parameters of their relationship with web scrapers by means of website architecture. Some websites have an application programming interface (API) that facilitates the gathering of information by users, including those who want to scrape their sites. One could reasonably assume that if a website has an API, its hosts agree to have their websites scraped within the framework set by the API. With websites that do not have an API, the situation is less clear, since the establishment of an API requires time and effort. Lack of an API may be due to, for example, limited human resources, and therefore, is not a reliable indicator of what resources the host wants to share. On the other hand, the position of site hosts regarding web scraping can be more reasonably inferred from the use of a variety of security methods to prevent web scraping, such as CAPTCHA (screening human from automated requests), blocking IP addresses and the use of honeypots (i.e. mechanisms to attract then block unauthorized users) and spider traps (i.e. mechanisms that cause the web crawling programs—also known as scripts—of unwanted users to crash or make an infinite number of requests).

Much of the current ethical discourse on web scraping is about relationships between web scrapers and web hosts, and the norms that should govern the harvesting of website information (Mitchell, 2015; Densmore, 2017; Krotov and Silva, 2018). A central question here is whether researchers using web scraping for health research purposes should hold themselves to moral standards more like those of biobank researchers than commercial web scrapers, and explicitly engage (and enter into agreements) with website hosts about the collection and use of the information they are gathering from them. If web scraping for health research purposes has the prospect for significant individual and social benefit, as well as exposure of identifiable persons to significant risk, one could argue that it is appropriate to enter into biobanking-like arrangements with hosts of websites from which data are scraped. Some emerging norms on big data research seem to point in this direction, though established practices that balance interests of web hosts and web scrapers are in their infancy (Zook *et al.*, 2017). But there are also strong reasons not to go far in this direction. Much of the data collected through web scraping are not specifically clinical or necessarily related to health at all. As with our study, the web-scraped information may only have significance for health when combined with other datasets, and therefore, requiring biobanking-like arrangements with webhosts is likely to be inappropriate in the vast majority of cases. The potential downsides of stronger formal regulation of health research web scraping are increased study costs and slow processes, which could undercut potential research benefits. To complicate matters further, forging agreements with webhosts may sometimes be simply impracticable, such as when thousands of websites are involved or websites have poor governance structures.

Respect for Recruited Participants and Study Communities

As Emanuel *et al.* note, clinical researchers can have obligations to research participants after they have been recruited and provided informed consent, and may also have obligations to the communities from which those participants come. For individuals, these can include confidentiality protections, the right to withdraw, compensation for study-related harms, continued post-research medical services and dissemination of research results. These obligations do not apply to all research studies, and some are not easily applicable to big data health research, including web scraping. For example, our participants do not provide informed consent, and they do not know they are in a study, so there is no way for them to exercise a right to withdraw. At the front end, one could, for example, notify those entering jail that the information to be placed on jail websites about them will be used as part of a study, unless they opt out. The burden of this approach on researchers and jail administrators would be extremely high. On the back end, if our data informed a D2C approach, the person contacted would have the right to reject the offer of health services, but that would be refusal of care rather than study withdrawal. Compensation for psychological, social and other harms due to inappropriate disclosure of personal information is relevant to web scraping and big data research, though in the US compensation structures are generally more developed in the commercial sector than in the research context. Some, like Vayena *et al.* (2015), propose the establishment of monitoring boards to devise compensation schemes for digital epidemiological research and surveillance. Such protections might be thought excessive for web scraping research using publicly accessible online data, but researchers should have *some* sort of contingency plans for data breaches, particularly when web scraped data are combined with nonpublicly accessible information. Responsible dissemination of results from big data health research is

complicated when individual consent is not involved and the research data were originally collected by other agencies (i.e., jails, courts, public health departments) for other purposes.

Engagement of communities in the research process has come to be regarded as scientifically and ethically desirable in many clinical and public health settings. In our study, meeting this requirement is challenging for two reasons. First, those who enter the North Carolina jails are research participants in the sense that at least some of them (in counties with jail websites) will thereby enter our database. While these individuals are the focus of the study and may be its potential future beneficiaries, the relationship between them and the researchers are much more remote than in much clinical research. The web scraping component of the research is indicative of this remoteness: it is an approach that does not involve interactions with individuals from which data are collected, but automated interactions between scripts developed by web scrapers and targeted websites. In addition, those individuals are not undergoing any sort of intervention, although the resultant database aims to inform future care efforts. Second, community engagement efforts related to studies like ours, using the internet and merging databases, demonstrate that the community to be engaged is unclear. Persons who have been in jail are a very diverse group that may not share needs and priorities, and it may be very challenging to find individuals who could be legitimately act as this group's representatives.

Beyond Emanuel *et al.*: Additional Ethical Considerations for Public Health Applications of Web Scraping Research

Is the Emanuel framework sufficient as a tool to help ethically guide and evaluate web scraping in public health research? As Ballantyne writes, big data research reveals that the boundary between research ethics and public health ethics is more a matter of emphasis and orientation than a hard line between incompatible frameworks (Ballantyne, 2019). We, therefore, suggest that some considerations drawn from public health ethics—more specifically as offered by Childress *et al.*, (2002) and Willison *et al.* (2014) should also be brought in, in cases where the goals, conduct and outcomes of such research overlaps with those of surveillance activities and public health interventions.

In their classic text, Childress *et al.* propose five 'justificatory conditions' in cases where pursuit of a public health good (such as improved HIV care and viral suppression among the incarcerated) may permissibly override or impinge on other moral values (such as autonomy and confidentiality). Examining these conditions may help to further articulate the ethics of web scraping in public health research.

The most basic condition, *effectiveness*, much like 'social value' in the Emanuel *et al.* framework, refers to the possibility that the activities in question will help improve public health. The issue here is whether research to develop a jail/court/HIV database to be used in enhanced surveillance and D2C is likely to improve health outcomes for those who are living with HIV and are jailed. Establishing an enhanced surveillance database would not be ethically justified if it was clear that it would never be used for those purposes. Whether research contributes to actual effectiveness will only be known gradually, as changes are made to systems and their effects are empirically evaluated over time.

Proportionality refers to public health benefits outweighing infringed moral considerations, such as using the data of individuals without their consent. This speaks to the importance of examining all benefits of public health-related research along with its negative effects. An interesting question is whether our research should be ethically evaluated with regard to how its database may be used in surveillance and D2C, since the latter is not within the researcher's control. However, if, for example, problems with the database could have an impact on persons—such as HIV-negative persons being more frequently contacted by public health services than they would otherwise—this seems to be a potential harm related to the translation of the research into practice. For translational research projects like this, researchers may have more obligation to work with public health authorities to anticipate and prevent such foreseeable harms.

Necessity refers in our case to whether a similarly useful database could have been created by without, for example, bypassing informed consent. Is that approach necessary? It is currently difficult to see a viable alternative in our case, but it is important to critically reassess this question over time. Commentators have noticed a rise in rhetoric about informed consent being superfluous in big data research, reflecting a sense that such socially beneficial research is unobtrusive, that asking for consent would lead to consent bias, that the data are already public, and that people would not mind anyway (Richterich, 2018). It is important to resist this trend by remaining skeptical about claims of consent (or other ways of respecting autonomy) being unnecessary or unfeasible, and whenever possible, to design ways of

pursuing public health goals compatible with individual autonomy and community agreement.

Least infringement, applied to our study, means that researchers should collect and disclose only the kind and amount of information necessary for the goal of empowering public health services to identify jailed persons living with HIV and re-engage them in care. Web scrapers should be guided by this consideration when they develop their scripts.

Public justification refers to the responsibility to explain and justify to the relevant parties why certain infringements (such as not obtaining informed consent) were done when collecting data for public health purposes. Sweeney *et al.* argue that public discussion, including solicitation of input from stakeholders, is crucial when planning and implementing a D2C approach using surveillance data. Ideally, the most affected population (persons living with HIV and who have been jailed) should be part of such discussions. One could argue that public justification is more appropriate for public health officials than researchers. However, to the extent that research has informed the D2C approach, researchers too should explain and justify their contribution.

Finally, Willison *et al.*, in offering an ethical framework for public health evaluative projects, also mention social justice considerations. This dimension is not prominent in the Childress *et al.* justificatory conditions and is especially important in research (like ours) focused on those already vulnerable to structural injustice, inequalities in power and stigma. Research projects should consider the extent to which their aims, activities or results are likely to ameliorate, reinforce or worsen existing inequities, and what resources can and should be leveraged by researchers to mitigate disadvantages.

While there may not (yet) be an overarching ethical framework for guiding and evaluating empirical research that uses big data methods to inform public health practice, as we suggest above, we believe some of its fundamental ingredients can already be discerned in the existing research ethics and public health ethics literature. Like Willison *et al.*, we believe that ethics guidance for activities in the gray zone between research and public health practice requires building outward from fundamental research ethics principles to embrace considerations that include community and population interests.

Conclusion

Web scraping may seem like an innocuous activity as it involves the mechanical collection of publicly accessible data, much like the everyday task of searching for information via a search engine. However, when web scraping is used for public health research, there are ethical implications that may not be obvious at first sight. We provide an example of a study using web scraping to gather jail website data in order to develop an enhanced surveillance database that could contribute to a future D2C approach for persons living with HIV who have been incarcerated in jails. In examining our own study, we believe that the identified ethical concerns are not so great as to prohibit moving forward with the project. However, our ethical assessment will continue to evolve as the project advances. Among others, a key issue moving forward is to better understand the social value and costs of the D2C paradigm in the jail context. For example, the justification of not obtaining informed consent while linking personal data among jail, court and public health agencies as part of a D2C program depends on how successful researchers and public health authorities are in designing a system to the HIV-related health to meet the needs of people incarcerated in jails. Other considerations of the project are whether HIV confidentiality can be protected during periods of incarceration and whether relevant communities can be adequately engaged to provide input regarding project activities. We will continue to monitor these issues and elicit input from relevant stakeholders as we continually assess the appropriateness of our project.

As Big Data health research continues to grow, it is likely that studies such as ours that combine web scraping, surveillance and initiatives to improve patient care will become more commonplace. Considerations from both research ethics and public health ethics will be relevant for their guidance and evaluation.

Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R01AI129731. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This publication resulted (in part) from research supported by the University of North Carolina at Chapel Hill Center for AIDS Research (CFAR), an NIH funded program P30 AI050410.

Conflict of Interest

None declared.

References

Ballantyne, A. (2019). Adjusting the Focus: A Public Health Ethics Approach to Data Research. *Bioethics*, **33**, 357–366.

Barocas, S. and Nissenbaum, H. (2014). Big Data's End Run around Anonymity and Consent. In J. Lane, V. Stodden, S. Bender and H. Nissenbaum (eds), *Privacy, Big Data, and the Public Good*. Cambridge: Cambridge University Press, pp. 44–75.

Centers for Disease Control and Prevention (2018). *Understanding the HIV Care Continuum*, available from: https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf [accessed 20 February 2020].

Centers for Disease Control and Prevention (n.d.). *Data to Care*, available from: https://www.cdc.gov/hiv/effective-interventions/respond/data-to-care/index.html [accessed 20 February 2020].

Childress, J. F., Faden, R. R., Gaare, R. D., Gostin, L. O., Kahn, J., Bonnie, R. J., Kass, N. E., Mastroianni, A. C., Moreno, J. D., and Nieburg, P. (2002). Public Health Ethics: mapping the Terrain. *The Journal of Law, Medicine & Ethics : a Journal of the American Society of Law, Medicine & Ethics*, **30**, 170–178.

Chiolero, A. (2013). Big Data in Epidemiology: Too Big to Fail? *Epidemiology*, **24**, 938–939.

Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Nagalingeswaran, K., Hakim, J. G., Kumwenda, J., Grinsztejn, B., Pilotto, J. H. S., Godbole, S. V., Mehendale, S., Chariyalertsak, S., Santos, B. R., Mayer, K. H., Hoffman, I. F., Eshleman, S. H., Piwowar-Manning, E. M. T., Wang, L., Makhema, J., Mills, L. A., de Bruyn, G., Sanne, I., Eron, J., Gallant, J., Havlir, D., Swindells, S., Ribaudo, H., Elharrar, V., Burns, D., Taha, T. E., Nielsen-Saines, K., Celentano, D., Essex, M., and Fleming, T. R.; for the HPTN 052 Study Team (2011). Prevention of HIV-1 Infection with Early Antiretroviral Therapy. *New England Journal of Medicine*, **365**, 493–505.

Densmore, J. (2017). *Ethics in Web Scraping*, available from: https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01 [accessed 20 February 2020].

Emanuel, E. J., Wendler, D., and Grady, C. (2000). What Makes Clinical Research Ethical? *The Journal of the American Medical Association*, **283**, 2701–2711.

Emanuel, E. J., Wendler, D., Killen, J., and Grady, C. (2004). What Makes Clinical Research in Developing Countries Ethical? The Benchmarks of Ethical Research. *The Journal of Infectious Diseases*, **189**, 930–937.

Goggin, B. (2019). *Inside Facebook's Suicide Algorithm: Here's How the Company Is Using Artificial Intelligence to Predict your Mental State from your Posts*, available from: https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12 [accessed 20 February 2020].

Gold, Z. and Latonero, M. (2018). *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping*, available from: https://digitalcommons.law.uw.edu/wjlta/vol13/iss3/4/ [accessed 20 February 2020].

Gostin, L. O. (2007). Biomedical Research Involving Prisoners: Ethical Values and Legal Regulation. *The Journal of the American Medical Association*, **297**, 737–740.

Grady, C. (2015). Enduring and Emerging Challenges of Informed Consent. *New England Journal of Medicine*, **372**, 855–862.

HealthMap (2006). Computational Epidemiology Group, Informatics Program, Boston Children's Hospital. Available from: https://www.healthmap.org/en/ [accessed 20 February 2020].

Hsu, J. (2018). *The Strava Heat Map and the End of Secrets*, https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/ [accessed 20 February 2020].

Iroh, P. A., Mayo, H., and Nijhawan, A. E. (2015). The HIV Treatment Cascade before, during and after Incarceration: A Systematic Review and Data Synthesis. *American Journal of Public Health*, **105**, e5–e16.

Krotov, V. and Silva, L. (2018). *Legality and Ethics of Web Scraping*, available from: https://aisel.aisnet.org/amcis2018/DataScience/Presentations/17/ [accessed 20 February 2020].

Mitchell, R. (2015). *Web Scraping with Python. Collecting Data from the Modern Web*. Sebastopol, CA: O'Reilly Media, Inc.

Mooney, S. J., Westreich, D. J., and El-Sayed, A. M. (2015). Epidemiology in the Era of Big Data. *Epidemiology*, **26**, 390–394.

Mutenherwa, F., Wassenaar, D. R., and Oliveira, T. (2019). Ethical Issues Associated with HIV Phylogenetics in HIV Transmission Dynamics Research: A Review of the Literature Using the Emanuel Framework. *Developing World Bioethics*, **19**, 25–35.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy and the Integrity of Social Life*. Stanford, CA: Stanford University Press.

Pappas, S. (2016). *How Big Is the Internet, Really?* available from: https://www.livescience.com/54094-how-big-is-the-internet.html [accessed 20 February 2020].

Richterich, A. (2018). *The Big Data Agenda: Data Ethics and Critical Data Studies*. London: University of Westminster Press.

Spaulding, A. C., Seals, R. M., Page, M. J., Brzozowski, A. K., Rhodes, W., and Hammett, T. M. (2009). HIV/AIDS among Inmates of and Releases from US Correctional Facilities, 2006: Declining Share of Epidemic but Persistent Public Health Opportunity. *PLoS One*, **4**, e7558.

Sweeney, P., Gardner, L. I., Buchacz, K., Garland, P. M., Mugavero, M. J., Bosshart, J. T., Shouse, R. L., and Bertolli, J. (2013). Shifting the Paradigm: Using HIV Surveillance Data as a Foundation for Improving HIV Care and Preventing HIV Infection. *Millbank Quarterly*, **91**, 558–603.

Vayena, E. and Gasser, U. (2016). Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine. In B. Middelstadt and L. Floridi (eds), *The Ethics of Biomedical Big Data*. Cham, Switzerland: Springer International Publishing, pp. 17–39.

Vayena, E., Salathe, M., Madoff, L. C., and Brownstein, J. S. (2015). Ethical Challenges of Big Data in Public Health. *PLoS Computational Biology*, **11**, e1003904.

Wassenaar, D. and Mamotte, N. (2012). Ethical Issues and Ethics Reviews in Social Science Research. In M., Leach, M., Stevens, G. Lindsay, A. Ferrero and Y. Korkutv (eds), *Oxford Handbook of International Psychological Ethics*. Oxford: Oxford University Press, pp. 268–282.

Willison, D. J., Ondrusek, N., Dawson, A., Emerson, C., Ferris, L. E., Saginur, R., Sampson, H., and Upshur, R. (2014). What Makes Public Health Studies Ethical? Dissolving the Boundary between Research and Practice. *BMC Medical Ethics*, **15**, 61.

Zook, M., Barocas, S., boyd, d., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., and Pasquale, F. (2017). Ten Simple Rules for Responsible Big Data Research. *PLoS Computational Biology*, **13**, e1005399.