SSM - Population Health 18 (2022) 101047

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

Population health science as a unifying foundation for translational clinical and public health research

Mark R. Cullen [a,d,*], Michael Baiocchi [b], Lisa Chamberlain [c], Isabella Chu [a], Ralph I. Horwitz [e], Michelle Mello [f], Amy O'Hara [g], Sam Roosz [h]

[a] Center for Population Health Sciences, Stanford School of Medicine, Stanford, CA, USA
[b] Department of Epidemiology and Population Health, Stanford School of Medicine, Stanford, CA, USA
[c] Department of Pediatrics, Stanford School of Medicine, Stanford, CA, USA
[d] Retired, USA
[e] Department of Medicine, Lewis Katz School of Medicine, Temple University, USA
[f] Stanford Health Policy and the Department of Medicine, Stanford University School of Medicine, and Stanford Law School, and the Freeman Spogli Institute for International Studies, all in Stanford, CA, USA
[g] McCourt School of Public Policy, Georgetown University Washington, DC, USA
[h] Crescendo, San Francisco, CA, USA

A B S T R A C T

Separated both in academics and practice since the Rockefeller Foundation effort to "liberate" public health from perceived subservience to clinical medicine a century ago, research in public health and clinical medicine have evolved separately. Today, translational research in population health science offers a means of fostering their convergence, with potentially great benefit to both domains. Although evidence that the two fields need not and should not be entirely distinct in their methods and goals has been accumulating for over a decade, the prodigious efforts of biomedical and social sciences over the past year to address the COVID-19 pandemic has placed this unifying approach to translational research in both fields in a new light. Specifically, the coalescence of clinical and population-level strategies to control disease and novel uses of population-level data and tools in research relating to the pandemic have illuminated a promising future for translational research.

We exploit this unique window to re-examine how translational research is conducted and where it may be going. We first discuss the *transformation* that has transpired in the research firmament over the past two decades and the *opportunities* these changes afford. Next, we present some of the *challenges*—technical, cultural, legal, and ethical— that need attention if these opportunities are to be successfully exploited. Finally, we present some *recommendations* for addressing these challenges.

1. The transformation of translational research for public health and clinical medicine

1.1. The legacy of translational research

Translational research in clinical medicine has a long and distinguished history, reified in 2003 by the introduction of the NIH "Roadmap." (Zerhouni, 2006) With the overarching goal of finding new and better medical treatments for the gamut of diseases, the process has proceeded along the pathway depicted in Fig. 1.

Against an essential background of research aimed at understanding biologic mechanisms more broadly and developing tools to support research relevant to multiple disciplines (often referred to as "basic science"), translational scientists have focused on 1) describing the clinical characteristics of diseases, often facilitated by assembly of patient registries; 2) using these detailed observations as the foundation for development of animal models—or, more recently, *in vitro* systems, including organoids, derived from animal or human tissues—which become the foundation for 3) explorations of the unique biology of each disease, and 4) the search for targets to disrupt the disease process. With maturity of this work and experiments suggesting such interventions might be beneficial, 5) medicinal chemists search for compounds or other moieties that might achieve that benefit in humans at an acceptable cost in terms of side effects and risks. After such an agent or device has been deemed ripe for testing, 6) trials commence starting with first-in-human tests to determine whether the effects in humans resemble that in animals and assess the dose-related adverse consequences. With that evidence in hand, the typical next step is 7) the conduct of a

* Corresponding author. 2776 Stirrup Way, Los Altos Hills, CA, 94022, USA.
E-mail address: phsdatacore@stanford.edu (M.R. Cullen).

https://doi.org/10.1016/j.ssmph.2022.101047
Received 27 October 2021; Received in revised form 31 January 2022; Accepted 13 February 2022
Available online 1 March 2022
2352-8273/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

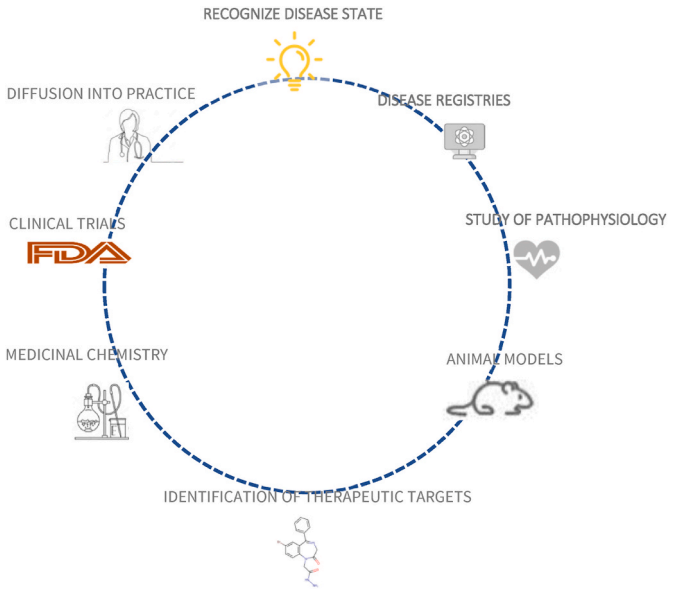

Fig. 1. The traditional pathway of discovery in clinical and translational medicine.

Translational research in *clinical medicine*, reified in 2003 by the introduction of the NIH "Roadmap" (Zerhouni, 2006), with the overarching goal of finding new and better treatments for the gamut of diseases, has proceeded along the pathway depicted in Fig. 1.

randomized controlled trial (RCT) of the new therapy compared to the standard of care or placebo.

By contradistinction, the search for evidence about public health interventions, such as nutritional supplements (e.g. Vitamin D in milk; fluoride in drinking water), environmental and occupational regulations, or policies to discourage harmful behaviors, historically has proceeded in a quite different way (Fig. 2).

Relying on vital records (births, deaths by cause, etc.), surveillance data, periodic community surveys, and assembly of large cohorts, epidemiologic research—cohort and case-control studies—became the primary tools for generating evidence to inform public health

Fig. 2. The traditional pathway of discovery in public health and epidemiology.

The search for evidence about public health interventions, such as nutritional supplements (e.g. Vitamin D in milk; fluoride in drinking water), environmental and occupational regulations, or the use of policies to discourage harmful behaviors, historically has proceeded in a quite different way than clinical research. This pathway is depicted in Fig. 2.

interventions. For example, observations of the long-term health of workers in various industries revealed the hazards of materials like benzene or asbestos; evidence of dose-related excess heart disease due to the public's exposure to fine particulate air pollution led to regulations to limit noxious exposures. Importantly, for many health problems, overlaps in efforts between clinical medicine and public health researchers occurred as analyses of different data sources suggested the utility of interventions at both the population and the individual level, e.g. vaccines.

Such projects engaged researchers from both fields, who brought different tools and perspectives to the table, for instance the effort to control HIV-AIDS (Piot & Quinn, 2013). Other times, differing perspectives generated some stress, highlighted recently by the tension arising over optimal application of initially scarce PCR tests for COVID-19: should they be used primarily for clinical diagnosis, or for surveillance to track spread of disease? But while some bridges have been built across researchers in clinical medicine and public health, until quite recently public health decisions have been grounded primarily on evidence from observational data, though learnings have often involved application of the underlying mechanisms involved, such as the pathways for disease transmission, biologic basis for risk, or mechanisms of action "borrowed" from medicine. Use of criteria such as those promulgated several decades ago by Bradford Hill and modified over time has rendered interpretation of observational research more consistent and palatable to those more confident of experimental approaches (Bradford Hill, 1965; Bradford Hill et al., 2020).

1.2. Sea changes in the last two decades

The past two decades have witnessed enormous advances in basic biology. Not only can we sequence an individual's genome at reasonable cost, we but we can measure the epigenome, transcriptome, metabolome and the scope and spectrum of the microbiome. Combining these advances in biology and those in data science, we can now scale the depth and breadth of our research to study large cohorts in which each individual's biology is characterized with petabytes of data, exemplified by the explosion of GWAS studies (Mills & Rahal, 2019). The concurrent expansion of biobanking has further afforded researchers the ability to quickly leverage new research modalities even for rarer patient populations (Ahadi et al., 2020).

The same vast expansion has occurred in the clinic as medical information has become digitized, essentially rendering complete health records part of the potential research quarry. Combining these two sources of information—clinical and biologic—has already yielded exceptional information about the role genes play in virtually every clinical condition (Tam et al., 2019). These analyses have also suggested that genes alone do *not* account alone for the fraction of disease believed to be heritable based on earlier studies (Boyce et al., 2020).

New opportunities in the era of big data are further enriched by the availability of vast, detailed, longitudinal data on environmental, social, physical, and behavioral factors that could link biology and *social factors* of populations with long-term outcomes (Rehkopf et al., 2016). Potential sources include not only the large *administrative* datasets held by government and private organizations, but also the troves of personal data collected *transactionally* on each of us every day as we use our phones, computers, credit cards, and customer loyalty program cards. Additionally, there are the rapidly growing repositories of "user-generated" data from fitness, health monitoring, and other apps; the geolocation data generated from geotracking technologies embedded in cellphones and smart watches; social media tracking of who we interact with, when, and where and the recordings made of our physical movement as we steer our car, move our computer mouse, or work the screens of our smartphones. In essence, we are all undergoing extensive psychometric testing all day, every day.

Deferring for now discussion of the myriad privacy concerns this raises, at least two previously unimaginable opportunities for

translational research become feasible. First, because these data are obtained in an ongoing fashion and many historical datasets have been digitized, following people and populations across the life course becomes possible beyond the older, painstaking strategy of long-term cohorts (Humphreys et al., 2018). Second, we can better link health to the many different ways we each—individually—lead our lives. Of course, this would be impossible but for concomitant developments in computer and data science. These huge leaps—e.g., the cloud and evolution of machine learning (ML)—elevate the analytical possibilities far beyond the traditional modeling methods upon which statisticians and epidemiologists have long relied (Chen et al., 2020a).

1.3. The opportunity for translational research

Numerous obstacles must be overcome in order to fully and responsibly realize the promise of the new data age for translational research. Before turning to these, we lay out the opportunity under the most favorable possible trajectory: all impediments can be overcome, and the resources needed to fulfill the promise can be garnered.

Fig. 3 visualizes a new paradigm for translational research in which the centerpiece is linkable, individual-level data derived from large populations. It depicts a research environment in which sources of biologic, clinical, physiologic, environmental, sociodemographic, transactional, and behavioral data are available for whole populations—serially—to facilitate a life course data panorama.

Many kinds of questions could be addressed within this data ecosystem. At the person level, data spanning the life course should allow linkage between conditions at early stages in life and later health. Each of the observed factors—medical, environmental or social—could be studied to generate hypotheses similar to the so-called "Barker hypothesis" that *in utero* exposure to food insecurity leads to later-life obesity (Almond & Currie, 2011). Every medical intervention could be traced forward into adulthood, indeed all the way to mortality. Complex aspects of life, such as work environment, social and neighborhood effects, life-long dietary exposures and habits, sleep patterns, and virtually every intervention that doctors and the health care system impose would become amenable to scrutiny in relation to virtually any short or long-term health outcome of interest. Where biologic and physiologic data on a sufficient sample are available and of high quality, not only the outcomes but the pathways between early causes and later outcomes might be elucidated.

Perhaps most exciting in this vision is the potential for "personalizing" our knowledge of these relationship based on our ability to predict likely responses to various therapeutic or preventive options. Failure of traditional studies to elucidate the optimal lifestyle suggest that "one size does not fit all" as we have begun to recognize for many drugs and medical treatments as well (Agarwal & Ioannidis, 2019; Markozannes et al., 2016). Average beneficial or harmful effects—the primary output of clinical trials and most observational studies—are exactly what the term implies: *average* effects across the population studied. Yet as we recognize the substantial diversity among us based on our unique biology and biography, what we want to discover are person- or person-type-*specific* treatment effects. This is most especially true in the realm of prevention, where presently most guidance regarding lifestyle, behavior, and environment is generic (one size fits all) (Arnett et al., 2019). This potential has already spawned speculation about the potential to personalize dietary recommendations (Topol, 2019).

2. The challenges

The data we anticipate will be central to the vision for Population Health Science as a unifying scheme for translational research are already being collected, some with the active consent and participation of the subjects, most passively. Collection of more or new data, per se, will *not* be critical. Making those data safely and securely available to the broad scientific community; creating new and refining existing computer and analytic tools; revolutionizing the culture and beliefs of the scientific and wider communities; and translating into practice the evidence they produce are the challenges that need attention (Leonelli, 2019). We approach these issues under three rubrics: technical challenges, cultural issues, and legal and ethical dilemmas.

2.1. Technical challenges

First, technical challenges must be surmounted to assure translational researchers access to relevant data and the ability to safely use and share them. Certain core principles have become axiomatic, frequently summarized by the acronym **FAIR** (Wilkinson et al., 2016). First, the data need to be **F**indable. In other words, there must be dataset search engines. The data need to be **A**ccessible, either to acquire or analyze on a suitable computational environment. To function as research data they must as well be structured in a format recognizable to each user, following a common data structure with shared data definitions, referred to as being **I**nteroperable.

Finally, the data need to be **R**e-useable by new investigators. This requires extensive documentation, generally referred to as "meta-data"

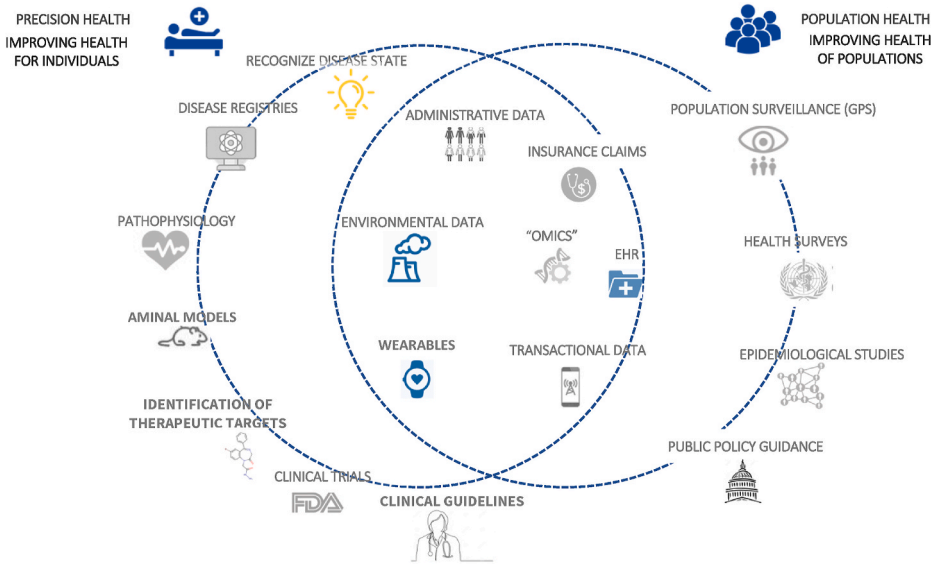

Fig. 3. A new paradigm for clinical translational research with large, linkable, individual-level datasets as the substrate.
Fig. 3 visualizes a new paradigm for translational research in which the centerpiece is linkable, individual-level data derived from large populations. It depicts a research environment in which sources of biologic, medical, physiologic, environmental, sociodemographic, transactional, and behavioral data are available—individually and longitudinally—for whole populations. Critically, while each byte of data is collected at a single point in time, those measures that change over time can be performed serially. Moreover, data from other points in time, including links to past administrative data, might be identified—imagine old tax and census records, birth and death certificates—and incorporated in such a way as to facilitate a life course data panorama.

which ensures all users know the assumptions, limitations, choices and concerns of the originators of the data. Only in this way can analyses be reproduced and replicated.

Government-held data meeting these standards, such as vital records, census, immunization registries, and all-payer claims data represent troves of unique researcher interest. While several federal agencies have collaborated with the Census Bureau to support the Federal Statistical Research Data Centers (FSRDC) (The Federal Statistical R, 2020) in making these granular data, including identifiable variables, available for research, the process remains cumbersome and expensive—all work must be done on site at one of the 31 centers—hence serving only a fraction of the potential research demand (Jarmin, 2021).

These requisites are also becoming a reality for the now more than 60 million individual datasets in wide use around the world that do **not** include proprietary, highly sensitive or otherwise personally identifiable information (PII). The challenge for health research is to extend such work to these sensitive datasets as well. While it is generally not necessary for a data *analyst* to have access to any directly identifying feature, the ability to merge datasets together—for example, to look at environmental measures in relation to health outcomes—demands that a data *manager* retains the means to link the files. In many cases precise geo-spatial information is sufficient. But even after the identifiers themselves have been stripped, the greater the number of variable fields on each subject, the easier re-identification becomes (Harron et al., 2017; Wirth et al., 2021).

Most often, at present, we evaluate requirements for datasets that do not include identifiers specifically enumerated in privacy laws by having institutional "experts" who render a judgment at the institutional level regarding privacy, hosting security and specific stipulations about user-access based on training and research credentials. The lack of standards for such classification is problematic, and the process fraught with potential for unwanted variation (McGraw & Mandl, 2021).

Of course, many datasets *are* high-risk. Efforts to mitigate their risks have proliferated, and are generally discussed under the concept of differential privacy, achieved by creating "synthetic" data—datasets with the same distributions as the original but in which no single file is unaltered (Boedihardjo et al., 2109). While possibly expedient in the short-term, such mitigations result in a sharp reduction in the long-term utility of the data, e.g. blurring geospatial coordinates, as have been done with many surveys to limit re-identification. Many partial technical fixes are evolving, institution by institution, from which hopefully will evolve a small number of best-in-show products that could be commonly adopted.

2.2. Cultural challenges

Four aspects of the present "culture" of the translational research community demand attention if the proposed vision for population health science can be achieved. First, there needs to be more universal expertise in the principles of this science among all those who contribute. Second, the present academic incentive structure for appointments, promotions and other rewards heavily rewards individual prowess and successful "labs," where teams are more likely to succeed going forward. Third, present deep biases in the relative value and utility of observational evidence, as opposed to that from randomized controlled trials will need to be reevaluated, and finally, the value of creation and sharing of critical research datasets must be more heavily rewarded. We discuss each in turn:

2.2.1. Training in population health science for translational researchers

Presently, research training in our academic medical centers (AMCs), and associated universities is well suited for the historic approach, illustrated in Fig. 1. Translational scientists, selected and promoted based on *individual* research prowess, garner the resources to embellish knowledge, typically in a narrow area, identifying collaborators with necessary ancillary skill sets or methods as needed. Core precepts, such as "prediction," "cause," or "standards of evidence," are relegated to specialists in those areas—biostatisticians and informaticians—and often adopted uncritically. Translational researchers in training, including MD/PhD candidates and post-doctoral fellows among others, have historically been taught far more about genetics and immunology than about applications of data science to clinical and population health problems, or causal inference; most are assigned to a wet lab very early on. Yet as we shift from the old paradigm, heavily dominated by development of animal or *in vitro* disease models in wet labs to the "information age" these imbalances and omissions will have to change.

Even as translational researchers increasingly make use of large datasets either as clinical research or in addition, they typically get more training in computer coding than in fundamental issues such as sampling strategy or analysis, or the difference between *predictors* of a health outcome and its (potentially treatable) *causes*. Invariably such investigators leverage the assistance of a colleague in informatics or statistics to apply the newest algorithms for machine learning or statistical testing, but revealingly, discussion sections of final reports focus more on putative biologic mechanisms —assuming the result is true— than critical assessment of the potential biases and limitations of study design (Giovannucci et al., 2008; Narod et al., 2019).

Two fundamental shortcomings stand out: 1) understanding the meaning of the "population" exploited for such research, and 2) confusion between *prediction* and *cause*. We elaborate on these two issues to illustrate the critical need for *all* translational researchers to be co-trained in data science broadly as, in the past, all have been trained in basic human biology.

So what is a "population," anyway? The term is now widely used to describe any large number of people with one or another feature in common, for instance: all 3 million people who have received care at a particular hospital; 12,000 patients with inflammatory bowel disease assembled from multiple patient registries; 10,000 participants in a public survey; or all children born in Denmark (Bengtsson et al., 2019).

No machine learning algorithm will discern, or alert an unprepared investigator to recognize, that inferences drawn from each of these "populations" will be different: some more representative than others of a larger population to which inferences may later be applied.

A parallel problem is the increasing, but potentially uncritical use of predictive models appearing in the biomedical literature (Luo et al., 2016). User-friendly statistical packages and their increasing availability make such analyses easy to conduct, but both users and consumers of resulting studies may lack understanding of what the models imply. There are many reasons one characteristic of a subject might "predict" a subsequent event: Fever in is a strong predictor of sepsis, but hardly a cause. Zip code is a strong predictor of excess hospital utilization but not a viable intervention target for individual patients (Chen et al., 2020b). Many of the strongest correlates of risk from Covid such as race and ethnicity have proved largely due to other, initially unmeasured factors such as essential occupation (Asfaw, 2021). Even strong correlates of outcome can be badly confounded, like serum beta-carotene, shown repeatedly to be a strong predictor of low cardiovascular and cancer risks, yet when tested as a supplement in an RCT it proved lethal (Omenn et al., 1996; Shekelle, Liu, Raynor, Lepper, & Maliza, 1981).

This is not to say we require knowledge of causal pathway or mechanism of action to optimally prevent or treat. Indeed, as we will discuss below, one of the putative benefits of RCTs as a source of evidence is that they typically *don't* require many assumptions about why one arm of a trial may prove more successful than another. But what we do want to have is evidence that modifying the single factor on which we intervene will (at least on average) improve an outcome of interest.

2.2.2. Translational science as a team sport

Many institutions recruit trainees and faculty for translational science using the principle of "best athlete"; the individuals most likely to achieve stellar personal success. Even where overall balance of faculty is

considered, the aim is not team building but breadth. And while most learn quickly the importance of cross-specialty input in preparing grants and papers—adding a statistician here, an economist or engineer there—the incentives for success are squarely on the "PI," who will be judged by the impact of their first and last authored papers in journals deemed of highest value to the PI's department.

Recognizing the limitations of such a structure, the notion of "team science" as integral to translational medicine research has achieved some cache over the past decade. Two distinct meanings have evolved. One is the concept of "broadly engaged team science" referring to the critical inclusion of all of the actors in late-stage translation, from trials to implementation. The focus of these teams is on inclusion of nonmedical professionals, patients, and members of stakeholder communities (Selker & Wilkins, 2017).

The second meaning refers to teams of scientists of very different skills and motivations who assemble to address problems that extend beyond the scope any one discipline. The COVID-19 pandemic offers a striking example: virologists, immunologists, geneticists, chemists, physicians, demographers, epidemiologists, computer scientists, mathematicians, economists, engineers ethicists, legal scholars, health behaviorists, health communications experts, and political scientists all play major parts. But this effort occurred under the extreme circumstances of a shared public and clinical health crisis, and has not been normative. To fully realize this conception of team science will require reorganization of existing research organization, with configurations of transdisciplinary teams, not PI labs "with consultants." Efforts to explore how such teams form, function, and survive in an academic universe not optimized around outputs that transcend narrow disciplinary norms has begun (Committee Toward an Open Science Enterprise, 2018; Stokols et al., 2008), but remains in its infancy.

2.2.3. *Hierarchies of evidence: critical acceptance of observational research*

The availability of rich population-level data could serve to markedly advance the efficiency and interpretability of many RCTs. For one thing, established "cohorts," with proper respect for privacy, can offer a ready-made template for trial recruitment; the preliminary observational analyses may further suggest an ideal sampling frame, pre-specifying subgroups for potential differential responses to the treatment in a prospective manner, and for addressing issues in generalization from recruits to larger clinical populations (Westreich et al., 2017). Once the study group has been selected, comparison with the larger observed population data could provide critical insight into how the study volunteers may, once selected, differ from the other potential subjects, impacting the interpretation of results and offering insight into the generalizability of the treatment effect measured.

But by far the biggest "gain" from the envisioned data-centered translational research universe will come from enhanced attention to the observational data themselves. And it is in this regard that the (presently) limited conversancy with the theory and practice of population health science among researchers in translational science has become rate-limiting. A strong belief has developed within the research community in which the evidence from randomized controlled trials is considered of materially greater value to decision making than observational data, however thoughtfully collected and analyzed, however well supported by ancillary scientific data (e.g. effects in animals) and however plentiful (i.e. replicated). There are very sound scientific reasons that experimental (RCT) data have achieved this preeminence, most notably that random assignment *is* the surest way to avoid the many "confounders" as noted above. Confounders include easily recognized and measurable relationships, e.g., smoking in examining the relationship asbestos and lung cancer (Klebe et al., 2019), or much harder to assess factors, like diet or stress. Most vexing of all for observational research are the myriad sources of "selection"—people and their doctors make choices for all sorts of reasons that themselves may be associated with different outcomes and generally difficult to directly observe.

Without disparaging the extraordinary benefit conferred on population level research by randomization, we propose to bring a pause to the often unbridled enthusiasm for the RCT as a research tool (Wang et al., 2015). Notwithstanding the biggest problem—many critical questions are not amenable to RCT for ethical or practical reasons—trials also have significant limitations (Deaton & Cartwright, 2018). Two concerns appear most salient. First is the belief that because neither measured nor unmeasured confounding factors can, by design, be correlated with treatment assignment in an RCT except by chance, no bias can creep in unless studies are poorly conducted. While this may be true for an 'instantaneous' assignment, where the entire treatment immediately follows randomization (such as surgery vs. stenting for CAD), non-random treatment "drift" occurs in trials that require longer treatment periods, as patients (non-randomly) drop out, take other interventions to treat side effects or even seek a supply of the active agent being tested. Use of the "Intention to Treat" approach provides a conservative solution for small drifts, but not larger divergences between assignment and treatment (Robins & Greenland, 1994). While strategies to adjust for these "late" biases have been developed, they involve approaches not unlike those used to address bias in observational studies (Hernán et al., 2013).

Perhaps the deeper limitation, though, relates to the output of RCTs: *Average* treatment effects (ATE, the absolute difference between prespecified outcomes in treated vs. control arms). Neither biographical not biologic attributes of subjects that may lead to heterogenous responses can be confidently estimated even for prespecified subgroups of interest because such subjects are typically too few compared to what could be observable in a study with real world data, such as post-marketing observations once an intervention is approved. As a result, RCTs have limited value in the effort to actualize "precision" or "personalized" medicine. Efforts to exploit large observational datasets as an alternative to exploring for heterogeneous effects have begun in earnest (Bodnar et al., 2020; Daoud & Johansson, 2019).

In the (Bayesian) scientific framework in which every study is premised on a foundation of prior beliefs, the notion that observational studies can only provide hypotheses for subsequent experiments is no longer tenable. Respectable observational studies demand a well justified conceptual framework in which all known or suspect causes, and their suspect inter-relationships are as prespecified as design of any RCT (Hernán & Robins, 2016; Robins, 1987). More than one hypothesis can be tested, with careful attention to statistical inference when multiple outcomes are considered simultaneously (VanderWeele et al., 2020; Vansteelandt & Dukes, 2020). In particular, where N is large and observations are rich, specific relationships between subject characteristics and outcomes of intervention can be tested. Methods have evolved over the last several decades to address biases, such as substitution for actual assignment by so-called "instrumental variables" (Marra & Radice, 2011; Rodu & Baiocchi, 2001) long used in economics—and dynamic marginal structural models for time varying covariates of concern (Robins, 1986). Newer methods are under exploration which would exploiting increasingly-available biologic "intermediate endpoints" (Athey et al., 2019). A novel contribution to the methodologic armamentarium, exploiting large data sets which have been genotyped, is "Mendelian randomization"—using the random assignment of measurable alleles as an instrumental variable to study environmental factors known to be directly impacted by that gene (such as variants of Apo E) (Smith, 2010). Myriad limitations of observational studies continue to be highlighted (Collins et al., 2020; Davey Smith & Phillips, 2020), while relevant concerns for interpretation of RCTS are typically ignored or downplayed.

In the end of the day both RCTs *and* observational studies are invaluable tools for translational research. No one doubts the importance of RCTs for definitive testing if new therapies confer more benefit on average than harm, as is apparent in the rush to treat COVID patients in the present pandemic. That said, much about the impact on health of human behaviors and exposures—including the role of medical

interventions and treatments—may best be learned from skilled analyses and inferential reasoning of increasingly rich person-level data.

2.2.4. Data sharing challenges

The transition to a research universe built on rich observational data cannot occur unless researchers can actually access these data. One persistent access barrier is the prevailing tendency of the research community to hoard academic assets once they are generated or obtained (Tenopir et al., 2011). Such hoarding not only detracts from the potential of population health science to generate new discoveries, but also hampers efforts to reproduce and replicate findings.

Presently, markets for research data appear to be absent from the academic marketplace. Researchers in all fields have historically exchanged data resources among small groups of collaborators, but in most, the practice of open data sharing has not evolved into a buoyant data market. Because the accumulation of academic reputation and credit is garnered primarily by high-impact publications (Nosek et al., 2012) and pioneering discoveries (Strevens, 2003), scientists are disincentivized to share research data. In addition, scientists in some fields face disincentives to act as customers of openly available data due to a perception that papers using such data are less impactful than primary (Wickham, 2019).

Over the last decade, the importance of data sharing for reproducibility, transparency, accelerated discovery, and collaboration has been recognized among stakeholders, including funding agencies, science agencies (Committee Toward an Open Science Enterprise, 2018) and academic journals (Alberts et al., 2015; Nosek et al., 2015). A growing number of journals encourage (and some mandate) making data supporting articles available (Vasilevsky et al., 2017).

Despite efforts to "open" data (and hence science), academic data exchanges to date lack an essential precondition for functional marketplaces: thickness (Roth, 2007). Too few scientists participate, because the data market is disconnected from the academic markets of scientific credit (Merton, 1973; Pierce et al., 2019) hiring and promotion (Moher et al., 2018). Shifting the culture to foster more data sharing will require research institutions to revise criteria for advancement to include production and sharing of high-value datasets. It will require journals to strengthen and enforce data-sharing requirements, and funders to better recognize the potential long-term value of expensive, laborious efforts to prepare high-quality data for use by others and facilitate its broad sharing.

2.3. Legal and ethical challenges

Of course, no matter how collaborative the research community becomes, much of the most relevant data arises not from academic research per se but from administrative data collection and curation as in health care and other daily business and government transactions. Access to—and responsible use of—the observational data crucial to the future of population health science hinges on our ability to address a spate of legal and ethical issues. Chief among these are concerns about data privacy and security and data use agreements (Ienca et al., 2018; Metcalf & Crawford, 2016; Mikal et al., 2016; Mittelstadt & Floridi, 2016; Rothstein, 2015; Stahl & Wright, 2018; Vayena et al., 2015).

2.3.1. Data privacy and security

Repeated reports of large-scale data security breaches has drawn attention to one harsh reality of the modern world described dispassionately above: we all live "under surveillance." Although reasonable people could debate the degree to which this is relatively benign or worrisome, it is clear that individuals have limited ability to control how much information is collected about them and how it is used. Some have curtailed their digital footprint by keeping Alexa and her friends out of their homes, limiting use of apps and online services, and setting their devices to strict "do not track/do not share" modes. But they do so at the expense of all of the services these apps and devices could provide, for free or at very low cost. Others concerned about data security and privacy have pushed for more regulation, in the form of strict limits to which data can be collected and how those data may be used. Yet others have pushed for a more market-driven solution, in which all personal data would reside legally and exclusively in the possession of the person *on* whom—not *by* whom—they are collected, leaving individuals in a position to sell or license some or all to bidders of their choice (Sonin et al., 2021).

The *privacy* of health information is only one among many related issues about big data that are currently under societal debate, but it has achieved particular salience. This stems in part from special protections given health data by law in most countries—so-called "health data exceptionalism." Also relevant is the widespread perception that health information is more intimate than information about other aspects of our lives, despite evidence that people may be even less forthcoming about, for example, their income (Tourangeau & Yan, 2007). In the realm of data *security*, health information actually has lower salience: it is less valuable to hackers than personal information that more directly facilitates lucrative crimes. It is relatively easy to identify potential harms caused by improper use or disclosure of health data. What is more difficult to measure is the opportunity cost of *failure* to use these data to their highest and fullest extent. As population health science advances three thorny privacy-related problems must be resolved.

First, the commercial-sector data ecosystem is too opaque. Most individuals have little or no awareness of the nature, scope, and value of the data trades they make every day when they use the internet and their devices. Put simply, "nearly everything done online involves trading personal information for things of value" (Cohen & Mello, 2018). Even information that is not, on its face, about health (for instance, income or neighborhood of residence) is useful—and increasingly used—to support modeling and inferences about health (Cohen & Mello, 2018). The world of commercially traded data is especially difficult to penetrate, but individuals may not even be aware of how their EHR information is used and passed on to third parties by their healthcare providers (Cohen & Mello, 2019). Nor are most individuals aware of the potential value of making these data—with proper privacy and security protections—available for observational research studies. Both sides of this issue should be elevated in the public consciousness so that those designing regulations and making everyday decisions about sharing their personal information can weigh the advantages and disadvantages in a more informed and deliberate fashion. For example, while few would doubt the public benefit that has accrued from the efforts to corral and analyze data on COVID-19, many would be disturbed to learn that no public tracking system in the US provides health investigators routine ongoing access to which individuals received the various vaccine preparations which would enable active adverse event case reporting reminiscent of a much earlier era in public health (Centers for Disease Control and Prevention, 2021).

Second, despite the potential for more informed decision making, there is reason for skepticism about perpetuating an information privacy regulatory scheme that leans on the notion of individual consent (McGraw & Mandl, 2021). Privacy laws in the US and abroad seek to ensure that individuals have an opportunity to authorize uses of their data to the maximum extent possible. The federal Health Information Privacy and Accountability Act (HIPAA), for example, provides that healthcare providers who collect identifiable health information electronically cannot disclose it to others, except for narrow purposes relating to treatment, healthcare operations, and public health reporting, unless patients authorize the disclosure. For research purposes, designated "Privacy Boards" (typically institutional review boards, doing double duty) can grant a waiver of this requirement, but only if several conditions attach, such as the impracticability of seeking patient authorization.

While the idea that patients should be able to control uses of their health information has strong intuitive appeal, it consistently falters upon execution (Canino, 2016; Kim, 2013; Meinel, 2016). Every patient

who has been asked to sign a HIPAA authorization form would agree that the process of reviewing and agreeing to these wordy, legalistic documents bears little resemblance to meaningful informed consent. Executing individual consent in the online context is even more farcical: research demonstrates that consumers do not read online privacy policies and end-user license agreements; moreover, even if they did, online service providers offer few or no alternatives to agreeing to the terms. In short, these permission-giving rituals are often hollow exercises that fail to effectuate the goal of meaningful consent and control over personal information. Yet, privacy law continues to rely upon them (California Office of the Attorney General, 2018; Wolford).

Third, health information privacy regulation relies on an outdated notion of "deidentified" data (McGraw & Mandl, 2021). When data are shared without personal identifiers attached, the transfer and use do not implicate the regulatory frameworks we have relied on for decades to protect individuals: federal human subjects research regulations and federal and state privacy laws (Cohen & Mello, 2018; Kaye, 2012). These laws date to a time in which reidentifying data that lacked personal identifiers was a practical impossibility, but advances in computing have greatly enhanced the technical feasibility of re-identification through data triangulation and hashing (Cohen & Mello, 2018; Kaye, 2012; Price & Cohen, 2019; Stead, 2017). "Deidentified" is increasingly recognized as a relative condition. Companies routinely approach health delivery systems to obtain "deidentified" patient health data (Farr, 2018). Although the datasets can be rendered "anonymized" based on deletion of PII, techniques to link the records with existing data are abundant (Harron et al., 2017; Wirth et al., 2021). Although such linkages have the potential to elucidate important and otherwise unanswerable questions about the relationships between social behaviors and health, the proposed arrangement would likely raise patients' hackles, but does not violate US law, suggesting the need for approaches to more thoughtfully weigh and adjudicate trade-offs.

Recently enacted privacy laws such as the CCPA (California Office of the Attorney General, 2018) and GDPR (Wolford) impose more stringent standards for considering a dataset deidentified, but do not decouple information privacy regulation from a determination about whether or not data are identifiable (McGraw & Mandl, 2021). Because current privacy laws push investigators to strip identifiers from datasets in order to reduce the risk that a privacy board or institutional review board will require them to seek individual consent for new uses of the data, they undercut potentially productive uses of data and limit the prospects for population health science. Consequently, scholars have suggested a need to reorient the law to "protect privacy while minimizing the cost to innovation" (Price & Cohen, 2019).

It is not even clear that the current regime addresses the concerns that animated its adoption. Although many consumers are concerned about potential consequences that may flow from wrongful disclosure or misuse of their identifiable personal information, for others the mere awareness that their personal data are accumulating on servers and in clouds of various organizations, including government, without their explicit consent, is *in of itself* a "harm" (Sonin et al., 2021). These individuals would *not* in general be willing to allow the use of their data even in putatively deidentified form. Indeed, some resist even participating in the U.S. Census and would not likely volunteer their data for any initiative without strict control over all present and potential future uses. While it is unclear what fraction of any population shares this perception, the moral weight of their argument offers a potent challenge to the open accessibility by researchers to population data.

2.3.2. Data use agreements

One factor impeding the efficient flow of data between data generators (including government agencies, private companies, and academic researchers) and secondary users in the research community is the length of time it takes to execute data use agreements, or DUAs (Major et al., 2020). These legal contracts, which spell out the rights and responsibilities of data generators and users and the remedies available to each party for breaches of the agreement, are negotiated on behalf of academic researchers by university administrators, and by legal counsel on behalf of nonacademic institutions. They are complex contracts, which augurs lengthy wait times for negotiation and execution (Microsoft, 2022; O'Hara, 2020). This can deter many investigators from seeking access to the best sources of data for their scientific question when inferior but more accessible sources appear to suffice (Mello et al., 2020). The exigencies of negotiation also may lead to compromises on DUA provisions that threaten researchers' academic freedom or ability to share data with others in the scientific community (Kanous & Brock, 2015).

Some problems contributing to delays in executing DUAs have ready solutions—for example, universities can increase staffing of the offices that handle them and create better portals for researchers to submit requests for a DUA (Mello et al., 2020). Some delays arise from persistent disagreements between the parties about particular provisions, however, e.g., data generators often demand data security architecture that is incommensurate the data risk profile or does not exist at universities (Mello et al., 2020; Saunders et al., 2015). For their part, universities insist on protecting researchers' rights to publish their research results, while many private companies are not acculturated to the importance of such freedom as a norm of academic science. On other matters, universities tend to resist making concessions with less justification: they may refuse to indemnify data generators in the event someone sues the data generator over some aspect of the research, for instance, although the actual risk of such a lawsuit is so low that it is not worth obstructing research over (Mello et al., 2020).

Perhaps the most fundamental obstacle to the timely execution of DUAs is that many private data generators, like academics, lack incentives to share data (Mello et al., 2020). Government and private companies generate an enormous amount of data of tantalizing research utility, but typically have no mandate or market incentive to allow researchers to use them. Possible exceptions include those in the business of healthcare itself, such as large public and private organizations that pay for healthcare or profit directly from it, such as pharmaceutical companies. Organizations in this sector have a positive incentive because of the economic value study results could produce, but also the reputational threat and potential liability that any breach or even public revelation of the research could present. For a great many other organizations, especially those in the digital-services business like Google and Facebook, managing, packaging and selling data for various kinds of analyses *is* a core business; there is no need to collaborate with academic researchers. Even when a company does perceive a business advantage from having a researcher answer a particular question, circumstances may change mid-course (O'Hara & Nelson, 2019). This creates an uncertainty hazard for the academic research enterprise, which relies on secure arrangements to assure completion of student projects and adherence to the rigid timelines of research grants and contracts.

3. Pathways forward

We have described a potentially exciting future for translational research, but also several challenges that must be surmounted to reach it. Next, we identify strategies for addressing the technical, cultural, legal, and ethical conundrums identified.

3.1. Addressing technical challenges

Simply put, the technical challenge is to achieve a state where qualified researchers working towards the broadest aims of translational clinical and public health research can avail data that are "**FAIR**" while at the same private and secure. While developments proceed on each component of this ideal condition—e.g., development of a standardized and automated instrument to assess re-identifiability of any data set, or machine learning tools that can rapidly "harmonize" data using differing data models—the ultimate ambition is to create safe research

ecosystems that incorporate these principles and are practicable in the research climate: It must be feasible, even attractive, for translational researchers in all settings to take full advantage. The trend has been towards development of "enclaves"—servers where the data of many relevant kinds are available with reasonable cost and effort, meet FAIR standards, and outputs of data off the enclave are surveilled for privacy risk.

Several specific efforts to achieve this merit special attention. First, the FSRDC model developed by the US Census Bureau, discussed above, is a useful exemplar for other governmental agencies to consider adopting (Jarmin, 2021; FSRDC, 2020). Importantly, not only are the number of available sites expanding but other governmental organizations in the US and around the world, are exploring smaller models, auguring a potentially rich role for the public sector in further development. (S special issue: data) The simple step of developing State All-Payer Claims Data troves for research now underway in many states, may be an important baby step in this direction (APCD Council, 2020).

Universities (Georgetown, 2022; Stanford, 2022) are also developing resources of this kind, if only as a stop gap to achieve data access to social and biomedical researchers on their own campuses. Nor is the private sector uninvolved: a consortium of private data vending and tech firms has been a leader in the provision of real-time, granular clinical and social data on the US population during the pandemic, providing a resource both to government and academic researchers when such data were not otherwise available (Datavant, 2022). Whether any of the non-governmental models is sustainable remains unproved, but momentum has been enhanced by the pandemic.

But to assure the future infrastructure of translational medicine will require that the major translational research funders—government and non-profit foundations—need to begin to enhance their investments and better coordinate their efforts. Presently, NSF, many of the NIH Institutes and Centers, and myriad foundations, global and domestic, have jumped into the fray to fund the underlying data science *methods*, for example novel approaches to differential privacy, or common data model development. But while advancing methods, including enhanced strategies for causal inference, are critical needs, they alone will not solve the broader infrastructure problem to achieve FAIR data, practicably available for all translational researchers.

3.2. Overcoming cultural barriers

The cultural barriers we have identified are possibly more formidable than the technical ones. Improving training in population health science; strengthening incentives for team science and data sharing; continuing to develop best practices for observational studies to build confidence that they belong higher up in the evidence hierarchy; and enhancing the incentives for data creation and sharing all must be pursued as part of the "long game" for translational research.

The most immediately actionable step is to begin enhancing the data and population science curriculum of medical and biomedical graduate students developing careers in translational research. For some training programs, such as MD/PhD programs, it might make sense to add such training as a prerequisite for admission rather than try to shoehorn it into the already crowded curriculum, or to encourage joint training programs with better established tracks in public health schools or programs. For others (e.g., post-doctoral research fellows in clinical departments), requirements and support for such training should become the norm. As these changes to training unfold, key allies of adding population science to the curriculum may be those waging the still-lonely fight to enhance scientific integrity and transparency.

Hiring, promotion, and recognition processes must further evolve to reward decisions to share rather than hoard, and to devote time and effort to creation of tools and resources that help others advance the field. The widely applied "impact" criterion could, for example, be interpreted broadly to include not just the ways in which scientists' research has changed thinking in the field, but also the ways resources they have created have enhanced the impact of others' research. Medical schools have already found ways to reward other material contributions faculty make, such as new technologies and intellectual property, patient referrals to trials, and the like; similar rewards could be developed for data-related contributions.

Two things need to happen first. There must be a simple way to count these contributions. The designation of standardized approaches to referencing datasets by journal editors, requirements that these be cited with every use of data, and establishment of standard ways of presenting them on CVs are crucial steps. Second, studies of "data markets" should be launched to establish the value to science of such contributions. The COVID-19 pandemic, which has brought unprecedented openness in the forms of preprinting and data sharing, may provide just such a natural experiment.

COVID-19 has also reinforced the value of team science: the problem is vast, multifaceted and not amenable to the solution a single lab could provide. Understanding the roles of host-factors, work, social behavior, and physical environment have been just as important to disease control as bringing vaccines to market at unprecedented speed. The value of team science will, in our view, win out over time without much additional deliberate intervention or promotion; our biggest immediate challenge is to train our workforce to adapt to and embrace this change. One idea to foster this is to re-examine the century-old split between schools of medicine and schools of public health, with an eye towards the emergence of "Schools of Health Science."

Finally, how can the culture of science be shifted to promote confidence in observational studies and disrupt established hierarchies of evidence? Clearly it will be important to further explore the limits of causal inference from observational data and develop of new methods and tools to address them. Research funders should earmark a pipeline of funding for this purpose. The ultimate objective is to enhance the utility of our growing trove of observational data and reduce reliance on RCTs to the settings in which they will add the greatest incremental value.

3.3. Approaches to the legal and ethical dilemmas

The legal and ethical dilemmas confronting translational science require a host of responses both short- and long-term. Increasing transparency around the "data trades" we make as consumers and improving public understanding of the actual and potential benefits of permitting responsible use of personal data, can and should begin now, while the pandemic experience is fresh in the public eye. Technical solutions to privacy problems should also continue to be pursued with vigor. New methods of safeguarding data and minimizing reidentification risks within and across datasets can help avoid wrenching decisions about whether to strip out useful but potentially identifying data fields from research datasets. They could also help build public trust.

A third short-term strategy for easing legal tensions is to promote standardization of the terms of data use in DUAs. Data generators and would-be recipients should not waste precious time haggling over points that should be non-negotiable (Mello et al., 2020). One promising development is the Federal Demonstration Partnership (FDP) project, in which 10 federal agencies and 90 research institutions are collaborating to identify ways of improving the efficiency of research; early efforts have focused on development of standardized DUA templates (Mello et al., 2020). In addition to supporting this approach, universities should increase staffing in the offices responsible for negotiating DUAs, recognizing that their workload has greatly expanded (Mello et al., 2020).

The long game for addressing legal and ethical tensions in observational research will be won by recentering our privacy protection regulatory regime so that it no longer balances precariously on the unstable pillars of individual consent and deidentification (McGraw & Mandl, 2021). In some contexts, such as prospective collection of observational data as part of a research study, it is both feasible and reasonable to require researchers to engage in an informed consent process with

prospective participants. But for secondary uses of information obtained for other purposes, whether online or in the physician's office, the hollow consent rituals that now dominate should be replaced, or at least joined, by deliberative, group consent approaches. As Cohen and Mello have argued in reference to secondary uses of EHR data, "Authorization that is individualized, upstream (i.e., obtained early), and typically one-and-done can be supplemented with governance that is group-based, downstream (i.e., obtained at the time of particular uses), and ongoing" (Cohen & Mello, 2019).

Two notable features of this approach is that the permission attaches not to the transfer of personal data but to specific uses; and that decisions are made by a multi-stakeholder committee that includes patient representatives but also experts in information technology and other fields who understand the potential privacy and security pitfalls associated with particular uses (Parasidis et al., 2019). Data ethicists have described other elements of a "systematic oversight approach" to data governance that would help undergird the oversight structure with more than the eroding concepts of individual consent and deidentification, and better balance the goals of protecting privacy and facilitating socially beneficial uses of personal data (McGraw & Mandl, 2021; Price & Cohen, 2019; Vayena & Blasimme, 2018).

Whatever permission-giving structures are adopted, it is clear that their remit going forward must include consideration of secondary uses of "deidentified" data. The assumption that uses of such data are risk free is simply outdated. Privacy laws should evolve to reflect this reality (McGraw & Mandl, 2021); their applicability or inapplicability to particular data transfers or uses should turn not on the presence of personal identifiers per se but on a broader assessment of risk and societal benefit. Further, these assessments should be made in a way that accounts for the rapid evolution of methods for triangulating and linking datasets.

In closing, we recognize the changes we propose are a tall order, and that resistance will continue to emerge from virtually every sphere. Rather than belabor the difficulty of that "squeeze," we choose to emphasize the "juice"—the enormous window of opportunity and potential benefit to every patient and individual who dreams of having decisions about their care guided by robust, highly specific evidence about patients like them while decisions around matters of public interest are guided by the best available evidence.

Author contributions

All authors shared in the design and approach to the paper. MRC was the lead author on this paper and responsible for the bulk of the writing. MB was made significant edits and contributed to sections on causality, LC made significant contributions to the conceptualization of the models in Figs. 1 and 2. IC, AO and SR contributed to the paper overall and made significant contributions to the discussion of data markets, legal and ethical issues around data sharing. RIH was made significant edits and wrote sections of the introduction and discussion, MM was largely responsible for the content sections on Data Use Agreements and Legal and Ethical Dilemmas.

Declaration of competing interest

The authors have declared that no conflict of interest exists.

Acknowledgements

We acknowledge the contributions of the Stanford Center for Population Health Sciences Data Core. The PHS Data Core is supported by a National Institutes of Health National Center for Advancing Translational Science Clinical and Translational Science Award (UL1 TR001085) and internal Stanford funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Agarwal, A., & Ioannidis, J. P. A. (2019). PREDIMED trial of mediterranean diet: Retracted, republished, still trusted? *BMJ, 7*, l341. https://doi.org/10.1136/bmj.l341. Published online February.

Ahadi, S., Zhou, W., Schüssler-Fiorenza Rose, S. M., et al. (2020). Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nat Med, 26*(1), 83–90. https://doi.org/10.1038/s41591-019-0719-5

Alberts, B., Cicerone, R. J., Fienberg, S. E., et al. (2015). Self-correction in science at work. *Science, 348*(6242), 1420–1422. https://doi.org/10.1126/science.aab3847

Almond, D., & Currie, J. (2011). Killing me softly: The fetal origins hypothesis. *The Journal of Economic Perspectives, 25*(3), 153–172. https://doi.org/10.1257/jep.25.3.153

APCD Council. (2020). *Interactive state report map*. Published https://www.apcdcouncil.org/state/map. (Accessed 27 July 2020).

Arnett, D. K., Blumenthal, R. S., Albert, M. A., et al. (2019). ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the American college of cardiology/American heart association task force on clinical practice guidelines, 2019 *Circulation, 140*(11). https://doi.org/10.1161/CIR.0000000000000678.

Asfaw, A. (2021). Racial disparity in potential occupational exposure to COVID-19. *J Racial Ethn Health Disparities*. https://doi.org/10.1007/s40615-021-01110-8. Published online August 5.

Athey, S., Chetty, R., Imbens, G., & Kang, H. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely* (p. w26463). National Bureau of Economic Research. https://doi.org/10.3386/w26463

Bengtsson, J., Dich, N., Rieckmann, A., & Hulvej Rod, N. (2019). Cohort profile: The Danish LIFE course (DANLIFE) cohort, a prospective register-based cohort of all children born in Denmark since 1980. *BMJ Open, 9*(9), Article e027217. https://doi.org/10.1136/bmjopen-2018-027217

Bodnar, L. M., Cartus, A. R., Kirkpatrick, S. I., et al. (2020). Machine learning as a strategy to account for dietary synergy: An illustration based on dietary intake and adverse pregnancy outcomes. *American Journal of Clinical Nutrition, 111*(6), 1235–1243. https://doi.org/10.1093/ajcn/nqaa027

Boedihardjo M, Strohmer T, Vershynin R. Private sampling: a noiseless approach for generating differentially private synthetic data. ArXiv210914839 Cs. Published online September 30, 2021. Accessed January 4, 2022. http://arxiv.org/abs/2109.14839.

Boyce, W. T., Sokolowski, M. B., & Robinson, G. E. (2020). Genes and environments, development and time. *Proceedings of the National Academy of Sciences, 117*(38), 23235–23241. https://doi.org/10.1073/pnas.2016710117

Bradford Hill, A. (1965). The environment and disease: Association or causation? *Sect Occup Med. Published online January, 14*, 295–300.

Bradford Hill, A., Armitage, P., Baiocchi, M., et al. (2020). Reprint of "the environment and disease: Association or causation?" with commentary. *Obs Stud, 6*(9), 1–65.

Canino, E. (2016). The electronic "sign-in-wrap" contract: Issues of notice and assent, the average internet user standard, and unconscionability. *UC Davis Law Review, 50*, 535–571.

Centers for Disease Control and Prevention. (2021). Vaccine adverse event reporting system (VAERS). Published November 2 https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vaers/index.html. (Accessed 16 January 2022).

Chen, S., Bergman, D., Miller, K., Kavanagh, A., Frownfelter, J., & Showalter, J. (2020b). Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care. *American Journal of Managed Care, 26*(1), 26–31. https://doi.org/10.37765/ajmc.2020.42142

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020a). Selecting critical features for data classification based on machine learning methods. *J Big Data, 7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Cohen, I. G., & Mello, M. M. (2018). HIPAA and protecting health information in the 21st century. *JAMA, 320*(3), 231. https://doi.org/10.1001/jama.2018.5630

Cohen, I. G., & Mello, M. M. (2019). Big data, big tech, and protecting patient privacy. *JAMA, 322*(12), 1141. https://doi.org/10.1001/jama.2019.11365

Collins, R., Bowman, L., Landray, M., & Peto, R. (2020). The magic of randomization versus the myth of real-world evidence. *New England Journal of Medicine, 382*(7), 674–678. https://doi.org/10.1056/NEJMsb1901642

Committee Toward an Open Science Enterprise. (2018). Open Science by Design: Realizing a Vision for 21st Century Research. In *Open science by design: Realizing a vision for 21st century research* (p. 25116). National Academies Press.

Daoud, A., & Johansson, F. (2019). *Estimating treatment heterogeneity of International Monetary Fund programs on child poverty with generalized random forest*. SocArXiv. https://doi.org/10.31235/osf.io/awfjt

Datavant COVID-19 Consortium. (2022). COVID-19 research database. Published January 1 https://covid19researchdatabase.org. (Accessed 16 January 2022).

Davey Smith, G., & Phillips, A. N. (2020). Correlation without a cause: An epidemiological odyssey. *International Journal of Epidemiology, 49*(1), 4–14. https://doi.org/10.1093/ije/dyaa016

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Farr C. Facebook sent a doctor on a secret mission to ask hospitals to share patient data. CNBC News. https://www.cnbc.com/2018/04/05/facebook-building-8-explored-data-sharing-agreement-with-hospitals.html. Published April 5, 2018.

The Federal Statistical Research Data Centers (FSRDC). Accessed November 11, 2020. https://www.census.gov/fsrdc.

Georgetown Massive Data Institute. Published January 3, 2022. Accessed January 7, 2022. https://mccourt.georgetown.edu/research/the-massive-data-institute/.

Giovannucci, E., Liu, Y., Hollis, B. W., & Rimm, E. B. (2008). 25-Hydroxyvitamin D and risk of myocardial infarction in men: A prospective study. *Archives of Internal Medicine, 168*(11), 1174. https://doi.org/10.1001/archinte.168.11.1174

Harron, K., Dibben, C., Boyd, J., et al. (2017). Challenges in administrative data linkage for research. *Big Data Soc, 4*(2). https://doi.org/10.1177/2053951717745678, 205395171774567.

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2013). Randomized trials analyzed as observational studies. *Annals of Internal Medicine*. https://doi.org/10.7326/0003-4819-159-8-201310150-00709. Published online September 10.

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial Is not available: Table1. *American Journal of Epidemiology, 183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Humphreys, J., Jameson, K., Cooper, C., & Dennison, E. (2018). Early-life predictors of future multi-morbidity: Results from the hertfordshire cohort. *Age and Ageing, 47*(3), 474–478. https://doi.org/10.1093/ageing/afy005

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review, 10. In G. Biemba (Ed.). *Plos one* (Vol. 13), Article e0204937. https://doi.org/10.1371/journal.pone.0204937.

IJPDS special issue: data centre profiles. International Journal of Population Data Science. 4(2). https://ijpds.org/issue/view/13.

Jarmin, R. (2021). Reflections on the successes and challenges of research data centers in Canada and the U.S.: Remarks at the CDRCN20 conference. *J Priv Confidentiality, 11*(1). https://doi.org/10.29012/jpc.765

Kanous, A., & Brock, E. (2015). *Contractual limitations on data sharing: Report prepared for ICPSR*. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/123016/ContractualLimitationsonDataSharing150411-1.pdf?sequence=1&isAllowed=y.

Kaye, J. (2012). The tension between data sharing and the protection of privacy in genomics research. *Annual Review of Genomics and Human Genetics, 13*(1), 415–431. https://doi.org/10.1146/annurev-genom-082410-101454

Kim, N. S. (2013). *Wrap contracts: Foundations and ramifications*. Oxford University Press.

Klebe, S., Leigh, J., Henderson, D. W., & Nurminen, M. (2019). Asbestos, smoking and lung cancer: An update. *International Journal of Environmental Research and Public Health, 17*(1), 258. https://doi.org/10.3390/ijerph17010258

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harv Data Sci Rev. Published online June, 22*. https://doi.org/10.1162/99608f92.17405bb6

Luo, W., Phung, D., Tran, T., et al. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research, 18*(12), e323. https://doi.org/10.2196/jmir.5870

Major, A., Cox, S. M., & Volchenboum, S. L. (2020). Using big data in pediatric oncology: Current applications and future directions. *Seminars in Oncology, 47*(1), 56–64. https://doi.org/10.1053/j.seminoncol.2020.02.006

Markozannes, G., Tzoulaki, I., Karli, D., et al. (2016). Diet, body size, physical activity and risk of prostate cancer: An umbrella review of the evidence. *Eur J Cancer, 69*, 61–69. https://doi.org/10.1016/j.ejca.2016.09.026

Marra, G., & Radice, R. (2011). A flexible instrumental variable approach. *Statistical Modelling, 11*(6), 581–603. https://doi.org/10.1177/1471082X1001100607

McGraw, D., & Mandl, K. D. (2021). Privacy protections to encourage use of health-relevant digital data in a learning health system. *Npj Digit Med, 4*(1), 2. https://doi.org/10.1038/s41746-020-00362-8

Meinel, M. (2016). Requiring mutual assent in the 21st century: How to modify wrap contracts to reflect consumer's reality. *North Carolina Journal of Law and Technology, 18*(5). Article 6.

Mello, M. M., Triantis, G., Stanton, R., Blumenkranz, E., & Studdert, D. M. (2020). Waiting for data: Barriers to executing data use agreements. *Science, 367*(6474), 150–152. https://doi.org/10.1126/science.aaz7028

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data Soc, 3*(1). https://doi.org/10.1177/2053951716650211, 205395171665021.

Microsoft. Removing barriers to data innovation. Microsoft Open Data. Accessed January 14, 2022. https://news.microsoft.com/wp-content/uploads/prod/sites/560/2019/07/Backgrounder-FAQ-Sheet_FinalV2.pdf.

Mikal, J., Hurst, S., & Conway, M. (2016). Ethical issues in using twitter for population-level depression monitoring: A qualitative study. *BMC Medical Ethics, 17*(1), 22. https://doi.org/10.1186/s12910-016-0105-5

Mills, M. C., & Rahal, C. (2019). A scientometric review of genome-wide association studies. *Commun Biol, 2*(1), 9. https://doi.org/10.1038/s42003-018-0261-x

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics, 22*(2), 303–341. https://doi.org/10.1007/s11948-015-9652-2

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLoS Biology, 16*(3), Article e2004089. https://doi.org/10.1371/journal.pbio.2004089

Narod, S. A., Huzarski, T., Jakubowska, A., et al. (2019). Serum selenium level and cancer risk: A nested case-control study. *Hereditary Cancer in Clinical Practice, 17*(1), 33. https://doi.org/10.1186/s13053-019-0131-7

Nosek, B. A., Alter, G., Banks, G. C., et al. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Office of the Attorney General, California. California Consumer Privacy Act (CCPA). Published January 1, 2018. https://oag.ca.gov/privacy/ccpa.

O'Hara, A. (2020). Model data use agreements: A practical guide. In *Handbook on using administrative data for research and evidence-based policy*. https://admindatahandbook.mit.edu/book/v1.0-rc4/dua.html#fn22 on 2022-01-16. (Accessed 16 January 2020).

O'Hara, A., & Nelson, J. (2019). *Evaluation of the social science one – social science research council – Facebook partnership*. https://hewlett.org/wp-content/uploads/2020/02/Facebook-Partnership-Final-Evaluation-Report.pdf. (Accessed 11 January 2022).

Omenn, G. S., Goodman, G. E., Thornquist, M. D., et al. (1996). Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *New England Journal of Medicine, 334*(18), 1150–1155. https://doi.org/10.1056/NEJM199605023341802

Parasidis, E., Pike, E., & McGraw, D. (2019). A Belmont Report for health data. *New England Journal of Medicine, 380*(16), 1493–1495. https://doi.org/10.1056/NEJMp1816373

Pierce, H. H., Dev, A., Statham, E., & Bierer, B. E. (2019). Credit data generators for data reuse. *Nature, 570*(7759), 30–32. https://doi.org/10.1038/d41586-019-01715-4

Piot, P., & Quinn, T. C. (2013). Response to the AIDS pandemic — a global health model. *New England Journal of Medicine, 368*(23), 2210–2218. https://doi.org/10.1056/NEJMra1201533

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nat Med, 25*(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Rehkopf, D. H., Domingue, B. W., & Cullen, M. R. (2016). The geographic distribution of genetic risk as compared to social risk for chronic diseases in the United States. *Biodemography and Social Biology, 62*(1), 126–142. https://doi.org/10.1080/19485565.2016.1141353

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling, 7*(9–12), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases, 40*, 139S–161S. https://doi.org/10.1016/S0021-9681(87)80018-8

Robins, J. M., & Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association, 89*(427), 737–749. https://doi.org/10.1080/01621459.1994.10476807

Rodu J, Baiocchi M. The principled prediction-problem ontology: when black box algorithms are (not) appropriate. ArXiv200107648 Stat. Published online April 3, 2020. Accessed November 11, 2020. http://arxiv.org/abs/2001.07648.

Roth, A. (2007). *What have we learned from market design* (p. w13530). National Bureau of Economic Research. https://doi.org/10.3386/w13530

Rothstein, M. A. (2015). Ethical issues in big data health research: Currents in contemporary bioethics. *Journal of Law Medicine & Ethics, 43*(2), 425–429. https://doi.org/10.1111/jlme.12258

Saunders, P. A., Wilhelm, E. E., Lee, S., Merkhofer, E., & Shoulson, I. (2015). Data sharing for public health research: A qualitative study of industry and academia. *Communication and Medicine, 11*(2), 179–187. https://doi.org/10.1558/cam.v11i2.18310

Selker, H. P., & Wilkins, C. H. (2017). From community engagement, to community-engaged research, to broadly engaged team science. *J Clin Transl Sci, 1*(1), 5–6. https://doi.org/10.1017/cts.2017.1

Shekelle, Richard B, Liu, Shuguey, Raynor, William J, Lepper, Mark, Maliza, Carol, et al. (1981). Dietary vitamin A and risk of cancer in the western electric study. *The Lancet, 318*(8257), 1185–1190. https://doi.org/10.1016/S0140-6736(81)91435-5

Smith, G. D. (2010). Mendelian randomization for strengthening causal inference in observational studies: Application to gene × environment interactions. *Perspectives on Psychological Science, 5*(5), 527–545. https://doi.org/10.1177/1745691610383505

Sonin, J., Becker, A. L., & Nipp, K. (2021). It's time for individuals — not doctors or companies — to own their health data. *STAT. November, 15*.

Stahl, B. C., & Wright, D. (2018). Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Secur Priv, 16*(3), 26–33. https://doi.org/10.1109/MSP.2018.2701164

Stanford Center for Population Health Sciences Data Portal. Published January 6, 2022. Accessed January 7, 2022. (phsdata.stanford.edu).

Stead W, National Committee on Vital and Health Statistics. Recommendations on de-identification of protected health information under HIPAA. Published online February 23, 2017. https://www.ncvhs.hhs.gov/wp-content/uploads/2013/12/2017-Ltr-Privacy-DeIdentification-Feb-23-Final-w-sig.pdf.

Stokols, D., Hall, K. L., Taylor, B. K., & Moser, R. P. (2008). The science of team science. *American Journal of Preventive Medicine, 35*(2), S77–S89. https://doi.org/10.1016/j.amepre.2008.05.002

Strevens, M. (2003). The role of the Priority Rule in science. *Journal of Philosophy, 100*(2), 55–79.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics, 20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tenopir, C., Allard, S., Douglass, K., et al. (2011). Data sharing by scientists: Practices and perceptions, 6. In C. Neylon (Ed.). *PLoS ONE* (Vol. 6), Article e21101. https://doi.org/10.1371/journal.pone.0021101.

Topol E. The A.I. Diet - Forget government-issued food pyramids. Let an algorithm tell you how to eat. New York Times. https://www.nytimes.com/2019/03/02/opinion/sunday/diet-artificial-intelligence-diabetes.html. Published March 2, 2019. Accessed November 11, 2020.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

VanderWeele, T. J., Mathur, M. B., & Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statistical Science, 35*(3), 437–466. https://doi.org/10.1214/19-STS728

Vansteelandt, S., & Dukes, O. (2020). Comment: On the potential for misuse of outcome-wide study designs, and ways to prevent it. *Statistical Science, 35*(3), 467–471. https://doi.org/10.1214/20-STS769

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ, 5*, Article e3208. https://doi.org/10.7717/peerj.3208

Vayena, E., & Blasimme, A. (2018). Health research with big data: Time for systemic oversight. *Journal of Law Medicine & Ethics, 46*(1), 119–129. https://doi.org/10.1177/1073110518766026

Vayena, E., Salathé, M., Madoff, L. C., & Brownstein, J. S. (2015). Ethical challenges of big data in public health. In P. E. Bourne (Ed.), *PLOS comput biol* (Vol. 11), Article e1003904. https://doi.org/10.1371/journal.pcbi.1003904, 2.

Wang, M. T. M., Bolland, M. J., & Grey, A. (2015). Reporting of limitations of observational research. *JAMA Internal Medicine, 175*(9), 1571. https://doi.org/10.1001/jamainternmed.2015.2147

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology, 186*(8), 1010–1014. https://doi.org/10.1093/aje/kwx164

Wickham, Rita J (2019). Secondary analysis research. *J Adv Pract Oncol, 10*(4). https://doi.org/10.6004/jadpro.2019.10.4.7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wirth, F. N., Meurers, T., Johns, M., & Prasser, F. (2021). Privacy-preserving data sharing infrastructures for medical research: Systematization and comparison. *BMC Medical Informatics and Decision Making, 21*(1), 242. https://doi.org/10.1186/s12911-021-01602-x

Wolford. (2021). *What is the GDPR, the EU's new data protection law?*. https://gdpr.eu/what-is-gdpr/. (Accessed 30 March 2021).

Zerhouni, E. A. (2006). Clinical research at a crossroads: The NIH roadmap. *Journal of Investigative Medicine, 54*(4), 171–173. https://doi.org/10.2310/6650.2006.X0016