nature aging

PERSPECTIVE

https://doi.org/10.1038/s43587-022-00277-x

The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging

Carol Brayne[1]✉ and Terrie E. Moffitt[2,3,4]

Large-scale volunteer databanks (LSVD) have emerged from the recognized value of cohorts, attracting substantial funding and promising great scientific value. A major focus is their size, with the implicit and sometimes explicit assumption that large size (thus power) creates generalizability. We contend that this is open to challenge. In the context of aging and age-related disease research, LSVD typically have limitations such as healthy volunteer, white ethnicity and high-education biases, and they omit early and late life stages critical for understanding aging. Their outputs are heavily focused on biomedical pathways of single chronic diseases. LSVD outputs increasingly dominate the funding and the publication landscapes. This Perspective discusses LSVD limitations and calls for more transparent reporting in LSVD research, as well as a greater reflection on the value of LSVD in relation to resources consumed. We invite funders and researchers to examine whether LSVD do actually contribute knowledge needed for our acute global health challenges including inequalities.

Cohort studies are research studies that collect data from a specific population of individuals over time to measure associations between risk factors (or exposures) and health outcomes. Many longitudinal cohort studies were initiated in the latter half of the last century to better understand aging trajectories and factors that influence health and longevity. These cohorts, from all life stages and across generations, have informed knowledge about origins of diseases including risk for and protection from poor health later in life. The remarkable expansion of biomedical science has brought detailed and ever-increasing genotyping and biological phenotyping to cohort studies, and many cohorts have incorporated such approaches, increasing the breadth of data gathered on individuals. Cohort studies' findings are of particular value because they can be interpreted in the context of their original population sampling. Care has often been taken with many of these cohorts that their sampling represents known populations.

Alongside these developments, improved biostatistical methods have driven recognition that very large numbers of participants are required to detect very small but potentially important biological pathways. Methods such as pattern detection, machine learning and ever-evolving computational tools are developed to 'mine' datasets. The creation of huge cohorts focused on mid-life and mainly based on large size emerged as a highly attractive new investment for researchers, funders, politicians and nations. We refer to these projects as LSVD. These involve recruiting research volunteers in numbers of hundreds of thousands to millions. Many LSVD have sprung up across the world this millennium, most particularly in high-income countries, including in the UK (UK Biobank), the US (All of US), Canada (CanPath), Germany (NAKO), Japan (Biobank Japan), Taiwan (Taiwan Biobank) and Finland (FinnGEN), among others[1]. Together with the usual extensive data collected from participating individuals by questionnaires, LSVD incorporate technologies to collect data in varied modalities, including multiple forms of omics and imaging. Many are linked to administrative records such as hospital admissions and primary care and social care records and are geocoded to contextual environmental data such as indices of air quality or socioeconomic deprivation. One of the most well known of these national investments is the UK Biobank (https://www.ukbiobank.ac.uk/), with half a million participants, 40–70 years old at baseline, supported by a £200 million investment. It is described on its website as a 'large-scale biomedical database and research resource, containing in-depth genetic and health information', 'a major contributor to the advancement of modern medicine and treatment' and 'a powerful resource for public health' (19 June 2022).

In this Perspective, we describe observations arising from discussions with colleagues and trainees, from trends in journals and media and from our experience as peer reviewers noting the exponential growth of LSVD publications. The promise of investments in LSVD is that they will provide definitive answers on the relationship of biological and sociodemographic factors with individual health outcomes that can then improve biomedical understanding and public health. As such, LSVD have attracted intense political, commercial, societal and scientific interest, leading to public investment on a scale approaching that for initiatives in defense, space exploration or atomic physics. Across countries, this LSVD approach has already reached several hundred million dollars collectively and is set to exceed billions in the future. Will LSVD yield findings that transform our understanding of disease processes and disease genesis? Many argue that they will and already have. But given that the pool of research funding is finite, others informally express concern that LSVD will, like a cuckoo in the epidemiological nest, consume an ever-increasing proportion of the health research monies that even high-income countries set aside. LSVD are set to be high-priced monocultures redolent of agribusiness, in which the same data soil is tilled by multiple harvesters, producing the crop of publications. There is an expressed concern that this monoculture creates an environment in which the diversity of our research approaches is being reduced (Fig. 1).

[1]Cambridge Public Health, University of Cambridge, Cambridge, UK. [2]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA. [3]Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK. [4]Promenta Centre, University of Oslo, Oslo, Norway. ✉e-mail: carol.brayne@medschl.cam.ac.uk

Fig. 1 | The potential creation of research monocultures. Will monocultures of data sources from a few gigantic fields of LSVD increasingly dominate the landscape of research?

We believe that now is a good time to step back with a reflection on the contribution of LSVD to the kind of knowledge generation required for the health challenges faced by global aging societies. In this Perspective, we lay out our areas of concern, and we call on funders and the research community to initiate a more systematic, independent review that would inform future decision making regarding LSVD investments. As experienced investigators who direct multi-decade and cross-generation smaller-scale cohorts and who have also published with LSVD and nationwide administrative linkage data, we will discuss our impressions of the limitations of LSVD and consider learnings and implications for future research, including the need for adequate scientific reporting, the distribution of investments, future costs, societal and environmental impacts and relevance in the context of health inequalities and sustainability. This is not a formal review; instead, we draw on examples when possible to illustrate specific comments.

A glossary of technical terms can be found in Box 1.

LVSD are not representative of the general population

An important part of the value of epidemiological studies depends on how closely their findings apply to the general population and therefore on how representative of that population their sample is. Compared to the general population, there are many reports from LSVD about how their participants differ[2–11]. Participants tend to be older, female, less urban, better educated, taller and white, have higher incomes, live in socioeconomically well-off areas, consult their doctors less often and require fewer medications. In other words, LSVD participants have less adversity and disadvantage across their life courses and therefore cannot be assumed to represent the general population and the diversity within populations. LSVD participants have been reported to have lower rates of most health risks and conditions. Examples from empirical publications include overweight and obesity, tobacco smoking, daily alcohol drinking, early menarche, asthma, dyslipidemia, dementias, neuroticism, schizophrenia, anxiety, depression, eating disorders and other mental health disorders, special educational needs, long-term illnesses, cancers, self-reported health conditions and all-cause mortality. This empirically documented 'healthy volunteer bias' compromises the generalizability of findings to more vulnerable populations who account for most of the burden of disease. And yet many LSVD publications implicitly lead to the conclusion that they can analyze life course adversity and that their findings generalize beyond their participants.

Many LSVD examine bias, and some have acknowledged at the outset that their volunteers were never intended to represent the population[2–4,8]. One often cited example, the UK Biobank, invited approximately 9 million UK residents, obtaining a 6% response rate of people who volunteered and became participants, a 6% that is known to be characterized by healthy volunteer bias[5]. One argument made in favor of recruiting healthier individuals at baseline is that some participants will in due course develop illnesses and therefore allow the study of health decline from baseline. However, this argument does not nullify concerns about representation, because, even when a condition develops in the healthier people in these studies, findings may not tell us much about the sort of health decline that accounts for the population's burden of disease. Perhaps this is best illustrated by a specific example: is it reasonable to assume that the course of diabetes that develops in a healthy, advantaged LSVD

Box 1 | Glossary

Attrition: attrition rates are values that indicate the rate of participant dropout in longitudinal studies.

Cohort study: a form of longitudinal study used in medicine and social science beginning with a group of individuals with a common defining characteristic, for example, a birth year.

Cohort effect: variations in characteristics among individuals who are defined by some shared temporal experience or common life experience or exposure, such as year of birth or exposure to leaded fuel.

Collider bias: if two unrelated characteristics both influence a third, such as the recruitment in a study, this can create the spurious appearance of associations.

Diagnostic criteria: set of agreed symptoms, signs and investigation results that make up a clinical diagnosis, defined and usually agreed upon by consensus among leading international experts.

Effect size: number measuring the strength of the relationship between two variables in a population or population sample.

Exposure: a variable with causal effect to be estimated including environmental and contextual factors, such as lived environments, medical conditions, treatments and genetics. Examples of exposures assessed by epidemiological studies are environmental and lifestyle factors, socioeconomic and working conditions, medical treatments and genetic traits. Exposures may be harmful, beneficial or a combination of both.

External validity: the validity of applying the conclusions of a scientific study outside the context of that study.

Large-scale volunteer database: databases made up of health information provided by hundreds of thousands of volunteers.

Omics: characterization and quantification of multiple biological measures that relate to structure, function and dynamics of an organism or organisms.

Healthy volunteer bias: volunteers for research have lower rates of disease and mortality than non-volunteers, a difference that can bias external validity of findings.

High-education bias: volunteers for research tend to have higher levels of educational attainment than non-volunteers.

Heritability: a measure of the genetic contribution to a phenotype in the population studied.

Incidence: the proportion of an at-risk population developing a given condition during a specified time period or aggregated person-years of observation.

Phenotype: an individual's observable states and traits, a combination in any given individual of genetic and environmental influences from pre-conception to end of life.

Prevalence: the proportion of a population with a defined condition or state at a specific time.

Provenance: 'The fact of coming from some particular source or quarter; origin, derivation' (ref. [55]).

Reverse causation: X and Y are associated but not in the way assumed because instead of X causing a change in Y, it is really the other way around: Y causes changes in X.

Survivor bias: survival bias is a type of sampling error or selection bias that occurs when individuals recruited for a study are those who lived past a certain age while those who died are ignored.

participant enrolled at the age of 65 years with a life expectancy of 100 years will generalize to the course of diabetes in a 65-year-old with diabetes onset earlier in life and who experienced lifelong major disadvantage? Probably not, but this is an empirical question. There is evidence of different natural histories depending on disadvantage for notionally the same condition (recorded in routine health record diagnoses, for example, see ref. [12]).

This rapid reflection on illness burden in whole populations quickly establishes that disease burdens in societies do not predominantly arise from the people taking part in LSVD. This is something that a check against epidemiological incidence patterns in true population samples can help to determine. A check would be particularly important for conditions such as cognition, dementia and frailty. We are concerned that the assumption of generalizability from LSVD to whole populations is a giant leap, not a small step. An illustration of this leap is the reporting of cognitive profiles in one LSVD, in which there may not only be participation bias but also gendered differences in such bias, rendering interpretation challenging[13]. We further discuss below LSVD biases and limitations and their implications for the generalizability of findings.

Age, life course, race and ethnicity. LSVD websites for databanks mention disparities, life course, public health and age-related outcomes. However, our informal but extensive scrutiny of what is predominant activity and what is emphasized in the promotion of these LSVD on websites and in media as well as in LSVD outputs suggests that instead there is a heavy focus on biomedical characterization of individual diseases and their molecular natural histories. In our opinion on extensive review of outputs, there is scant mention of quality of life or well being, even when there has been measurement of relevant variables. One illustration is the German Cohort Study (NAKO), in which the emphasis is as stated above but with broader intention including health economic analysis. Given the response rate of under 20% for NAKO (https://nako.de/, visited 19 June 2022) including challenges in migrant population representation and retention, there is clearly a jeopardy in assuming that analyses from this study will really represent population need in Germany[14].

An important consideration in the context of adequate population representation is age. Although LSVD's recruited age groups do vary and some include early adulthood, there is an emphasis on participants in mid-life to early late life. Although some LSVD remotely track mortality and care service engagement, most LSVD miss the stages at the end of life. Most also miss the early-life stages, despite the fact that longitudinal cohort studies have consistently established the vital importance of early-life stages for adult health. Early-life exposures and characteristics such as childhood adversity, lead exposure, childhood cognitive ability, childhood self-control, adolescent tobacco smoking and educational attainment are able to predict older-adult morbidity and mortality outcomes decades later[15]. When such key data are excluded, this creates challenges for the interpretation and generalizability of findings. As an example, LSVD may gather participants' retrospective reports of childhood, but retrospective measures of early-life adversity, in which adult respondents recall their childhoods, are known to identify

different people than do valid prospective adversity measures and to yield different findings[16]. This concern is not only limited to LSVD but applies to studies with later-age recruitment. The scale of such variation of findings according to sampling age is, we believe, underexplored.

Compared to the general population, fewer databank volunteers are black or brown people or those from diverse cultures. A few LSVD do aim to improve representation of such understudied ethnic ancestry or cultural groups. As one example, the American All of Us databank, aiming for 1 million participants, has enrolled about 275,000 volunteers as core participants since 2018 and reported that 80% of them belong to understudied groups[17]. However, there are concerns in the research community that individuals from minority ethnic groups who volunteer are not necessarily representative of their ethnic populations. The potential for differential response rates by ethnicity has not been documented systematically, but it is widely known that it is difficult to recruit minority ethnic participants and that many ethnic groups are reluctant to engage with medical research. Exacerbated healthy volunteer bias seems likely to characterize the members of understudied groups who are willing to volunteer and be retained[14]. Moreover, because of higher mortality rates at all life stages in minority ethnic populations, survivor bias also seems likely to disproportionately characterize LSVD participants from under-represented ethnic groups. Survivor bias occurs when individuals recruited for a study are those who lived past a certain age and those who already died are ignored. As a consequence, those participants who do take part in LSVD may be even less representative of their peers than participants from mainstream groups. There is a danger that adding extra weight to the data from ethnic participants who are healthier and more affluent than their ethnic group in general will exacerbate data bias rather than improve it as claimed.

Diseases, illnesses and disorders can present differently in different socioeconomic and age groups as noted above[12]. Thus, LSVD findings may provide value only for certain groups within the population. As an example, most dementia and severe cognitive impairment in high-income countries occurs in the older old and those close to death[18,19], whereas onset occurs earlier in people from disadvantaged or minority groups[20]. As the old, the poor, those from minorities and the vulnerable are under-represented in LVSD, findings on conditions such as dementia and severe cognitive impairment are not generalizable to these groups. A key illustration here is that the earlier-onset dementias in LSVD participants who tend to be under the age of 80 years (only a minority of dementia incident cases in high-income countries) will be associated with clear-cut pathologies such as signs of Alzheimer's disease. By contrast, by far the most dementia cases in the population onset after the age of 80 years and are less clear-cut, mixed pathology in nature[21–23].

Estimation of disease occurrence and risk. As discussed above, it is clear that LSVD are prone to the healthy volunteer bias that emerges because people who volunteer to join LSVD tend to come from more advantaged sectors of society and are healthier than a randomly selected sample of the population. This initial recruitment bias is likely further exacerbated because participants who continue with the databank over time tend to be even healthier than initial joiners who later drop out or provide incomplete data. Especially relevant for aging research, healthy volunteer bias is even worse among the LSVD's older-adult members because unhealthy adults of the same birth generations may have already died before the databank recruited its participants or will die before follow-up (survivor bias). Whether neuroimaging findings can truly provide age measurement as claimed in one output from the German Cohort Study requires scrutiny of just who participated in this intensive phase of investigation, given the many exclusion criteria for neuroimaging[24].

Healthy volunteers in LSVD are also genetically different from the general population[6,7,10]. In genome-wide analysis, single-nucleotide polymorphism-based heritability explained a fifth to a third of ongoing databank participation[10]. People carrying genes known to increase risk for tobacco smoking, overweight, neuroticism, schizophrenia, attention-deficit hyperactivity disorder and depression were less likely to provide data to LSVD[6,7]. People carrying genes associated with higher educational attainment and better health are more likely to take part in LSVD[25].

One of the consequences of the healthy volunteer bias is that LSVD are unsuitable for estimating the population's prevalence of a health condition. We have observed that scientific findings about prevalence are unfortunately often reported and accepted at face value. This observation is based on our scrutiny of publications arising from databanks, including a specific search of the UK Biobank website for the term 'prevalence', and is illustrated in these examples[26,27]. From an epidemiological and public health perspective, however, prevalence or incidence estimates must be obtained from a population of known provenance and claims about generalizability can only be supported by demonstrating that the population studied is indeed representative of the relevant population.

Some supporters of LSVD have stated that a sample's lack of representativeness should not be regarded as a scientific limitation and have reported that healthy volunteer bias should not reduce science consumers' trust in their estimates of associations between past exposures, diseases and future prognostic outcomes[28–33]. It has also been argued that LSVD show less dropout and attrition than would be expected in samples that began as population representative. It has even been asserted that representativeness is over-rated and is not a reasonable research aim[33]. These arguments should be open to careful scrutiny including simulations and biostatistical testing because many experts disagree with these reassurances that it is safe to make inferences from findings based on LSVD[7,25,34–38]. Many experts advise against estimating the co-occurrence of diseases in a databank with healthy volunteer bias because two diseases might seem to be associated when they are not. For example, if two characteristics both influence volunteering for an LSVD, this can create the spurious appearance of associations in the data, known as collider bias[39]. Volunteers also tend to come from geographical areas that have different prevalence of diseases compared to the population, which can generate spurious associations. Research on coronavirus disease 2019 (COVID-19) provides an illustration of these concerns. UK Biobank members who were tested for COVID-19 showed even more bias with regard to genetic, behavioral, health and demographic characteristics than the rest of UK Biobank participants, and authors cautioned against drawing conclusions about COVID-19 disease from these participants, because of the likelihood of collider bias[40].

Experts caution that the size of associations estimated in healthy volunteers is often inaccurate, noting that healthy volunteer bias can not only inflate associations but sometimes can deflate associations toward the null[25,38]. Shrunken effect sizes are a problem for health policy because they could conceal true causes of disease from sight. Exaggerated effect sizes, on the other hand, pose a different problem because they raise expectations that similarly large effects will apply later, in real-world diverse populations. Such expectations are often disappointed, but disappointment usually only ensues after an inflated finding has led to funding of costly prevention schemes and clinical trials. In one well-known example, a large volunteer study of nurses reported that hormone-replacement therapy was associated with a halved risk of heart disease. This finding prompted a large randomized clinical trial of hormone therapy that subsequently, after great expense, revealed that there was no real association between hormone replacement and heart disease[41]. The original finding was inflated in part by healthy volunteer bias because participants who used hormone-replacement therapy also tended to

Table 1 | Recommendations for best practices and transparent reporting in studies using LSVD

Section	Recommendations
Title	Include the word 'volunteer' where the title states the design/method.
Abstract	State that sources of bias are acknowledged in the article.
	Mention key data descriptors, such as response rate, age, sex or ethnic group.
	Refrain from making claims about generalizability in the abstract.
Methods	Report the rate of participant recruitment to the databank.
	Report rates of follow-up attrition relevant to analyses in the manuscript.
	Give the primary reasons for non-participation at recruitment and follow-up.
	Report rates of missing data for key variables in the manuscript and explain how missing data were addressed in analyses.
	Report comparisons between the databank and its general population.
	Cite prior publications that have investigated healthy volunteer bias in the databank.
	Use structured reporting guidelines such as STROBE and include corresponding checklists as part of the supplementary files.
Results	Provide a flow diagram showing numbers of individuals at each stage of the manuscript's analyses: for example, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, passing exclusionary criteria for analyses and analyzed.
	Describe attrition effects by including empirical tests comparing databank participants in the manuscript to databank participants not in the manuscript on key variables from the time of initial recruitment.
	Analyze a more population-representative subsample drawn from the larger databank to check whether effect sizes from the whole databank differ or remain the same. Report findings.
	If possible, test for replication of the manuscript's key findings using a different cohort that is more population representative than the databank. Report findings or state that this test was not feasible, explaining why.
	Consider using approaches such as inverse probability of attrition weighting[48] or survivor average causal effect[49].
Discussion	Include a paragraph dedicated to the discussion of the limitations of the findings, explaining differential response and sources of potential bias or imprecision, such as age range, birth year, nationality or ethnicity.
	Discuss potential biases in terms of both direction and magnitude.
	Discuss the generalizability (external validity) of the study results. Specify groups omitted to whom findings may not apply.
	Explain to readers in plain language that healthy volunteer bias calls for caution when generalizing findings beyond the databank.
	Discuss the possibility of collider bias, if relevant.
	Explicitly mention that observational studies do not establish causal relationships and discuss the possibility of reverse causality.

have healthy lifestyle behaviors known to reduce the risk of future heart disease.

The need for more transparency in reporting

The best-funded LSVD have created quasi-industrial processes to ensure quality control and standardization in data collection. While some of the research is carried out by scientists directly involved in these databanks, the data resources generated are by design 'open' and meant to be mined by numerous other scientists who were not involved in the data-collection efforts, at their 'Researcher Workbenches' (for example, ref. [42]). This industrialization of data capture and the open-access mandates associated with LSVD data have some undeniable value. However, it also means that data analysis is detached from knowledge of participants within their communities, which increases the risk that the provenance, population terms and deep knowledge of variables in such datasets will not be fully understood by data users or by consumers of LSVD findings.

The enormous numbers of participants in LSVD are impressive and, as a consequence, the media, politicians, the public and even scientists may assume that any finding that has been derived from a dataset based on so many people must be both true and important. Neither are necessarily the case. Warnings about the risks of healthy volunteer bias have been too frequently ignored or dismissed by researcher–analysts of LSVD or by science consumers using the outputs of LSVD. Many do not know about healthy volunteer bias and its attendant risks. Partially informed consumers place (understandably) too much confidence in findings[35,41,43]. Proponents of LSVD are partly to blame for less-than-transparent communication. For example, published statements reassure readers that the large size of the UK Biobank in itself guarantees that findings can be generalized to the full population with statements such as 'although UK Biobank is not suitable for deriving generalizable disease prevalence and incidence rates, *its large size* and heterogeneity of exposure measures provide valid scientific inferences of associations between exposures and health conditions that are generalizable to other populations' (ref. [5]; our emphasis).

It has been noted by some that many researchers who use LSVD data are not sufficiently transparent when communicating their findings[29,44,45]. Although there are some excellent cohort-specific reflections on strengths and limitations and analytical approaches to try to mitigate challenges, these are not universal nor, in our view, can the methods entirely address the limitations[46]. Some LSVD researchers using the data do make extra efforts to examine external validity of the data, with outcomes that are sometimes reassuring and sometimes not[5,17,26]. However, many researchers focus on size, precision and power, ignoring healthy volunteer bias, perhaps not understanding it themselves. Sometimes an apparently high response rate is quoted from the most recent stage of an LSVD, without mentioning that the initial stages had a tiny response, thereby misleading those reading into thinking that the response rate indicates generalizability[24]. It is also unfortunately rare for publications from LSVD to report rates of recruitment and rates of follow-up attrition, present empirical comparisons between sample and population, check effect sizes on a 'quasi' population-representative

subsample drawn from the dataset or even cite prior publications that report on the sample's healthy volunteer bias. These are the basics of good science communication, and they are easy to do but easy to forget or disregard. Not enough LSVD papers explain to readers in plain language that healthy volunteer bias calls for caution when generalizing findings beyond the sample. One reason for the transparency and reporting problems is that perhaps relatively few LSVD users are trained in population-based research. When published articles do take the trouble to be forthright about uncertainty, the findings are generally more trusted and become more influential, a lesson that we can learn from the history of climate science.

In our experience, many peer reviewers and journal editors who do not appear to be skilled in population research allow authors to imply inference of validity and generalizability beyond LSVD participants. Often science 'consumers' including media and politicians amplify these claims. Checklists, such as STROBE (https://www.strobe-statement.org/) or STROND[47], have been developed to use when generalizability to populations is being suggested. These have still not been sufficiently widely adopted by authors or journals. Further approaches to address the biases noted have been developed including approaches for application to studies focused on aging[48,49]. To help achieve more widespread best practices and transparency in the scientific reporting of LSVD studies, we list in Table 1 a number of recommendations. As an additional possible step to improve reporting, we also propose that users are made aware of the caveats that we raised here and should be provided by LSVD data distributors with examples of analyses when appropriate phrasing and sensitivity analyses or validated analytical approaches, such as weighting, are available[46]. This should then become the norm, with publications being expected to demonstrate up front how results might be limited or biased and in what direction.

Learnings from small cohort studies

The considerable limitations of cohort studies have been spelt out over many decades[50], including limitations of statistical power and attrition biases, but there is no doubt that these studies have reached deeply into people's lives across decades and indeed generations. Smaller cohort studies are often geographically or community based and are critically dependent on building lasting relationships of trust with their participants to ensure long-term participation. Both authors have been involved in long-standing studies in specific communities over decades, including early and late life. Responses of members of a 5-decade legacy cohort study were compared against responses of research-naive age mates, revealing that cohort members are far more willing to report frankly about illicit activities including drug abuse, risky sexual behaviors, intimate partner violence, child maltreatment and illegal income sources, thereby generating better data quality (T.E.M., personal communication). This data quality could not have been achieved had the research team not earned the cohort members' trust through strict confidentiality and data security. A high-response, geographically defined cohort of people aged 75 years from the mid-1980s that was followed up to death has informed knowledge of late-life transitions including cognition, frailty and quality of end of life. Cohort studies have also recorded the more contextual and multidimensional life experience of their participants and families (https://www.cc75c.group.cam.ac.uk). These include birth cohorts, older person's cohorts and cohorts focusing on disadvantaged communities.

Contemporary cohort studies that have evaluated their capacity to recruit from unhealthy vulnerable populations, especially participants of advanced age, have shown that this recruitment is intense and can be difficult, requiring dedication and major investments of finance, skills and time to build enduring trust and long-term goodwill between participants and researchers. The financial investment required is not judged in terms of the value of the contribution, although it pales when compared to the sums consumed by LSVD. The need for community trust and engagement to facilitate research of relevance to people's lives in their contexts is starting to be recognized by funders such as the UK's National Institutes of Health Research, which is increasing investment into local systems for research approaches embedded within communities.

Some may argue that most small-scale cohort studies that start out with population-representative samples also develop healthy volunteer bias at subsequent follow-ups after selective attrition accumulates. However, smaller-scale cohorts that recruit young people have the key capacity to link back to their original populations to compare retained participants against the base population in a systematic formal analysis of attrition bias. Such attrition analyses reveal where the biases are and what the biases' impact might be. Given their reliance on volunteer publicity approaches to recruitment, LSVD are not able to provide this vital information.

How can our many concerns be addressed? One option is to ensure that health science avoids the danger of all eggs being in one basket. Funders and researchers should consider the diversity of research designs relative to the diversity of the health challenges that societies face. 'Boutique' studies that are smaller could provide some of the anchoring required for the LSVD data. Analytical approaches such as weighting can help, but these cannot create the population groupings that simply are not there in LSVD. A careful reflection is required about the true place, strengths and limitations of the LSVD data themselves. Some of the contemporary cohorts have worked hard to ensure the link to populations. Examples include the Irish national aging study TILDA as a nationally recruited, locally grounded cohort with an emphasis on multidisciplinary collaborations that explore lives as they are lived in their community[51] (https://tilda.tcd.ie/). The Canadian Longitudinal Study on Aging, although experiencing low response rates, also has a regional engagement approach and transparent attention to respondent derivation[52]. Other data projects are larger but retain a successful focus on population representation and therefore better likelihood of generalization. These include the US Health and Retirement Study and large studies based on health care provision to defined populations such as Kaiser Permanente or the Veterans and Health Maintenance Organisations (albeit with the limitations of who is in the Health Maintenance Organisations). The Dunedin Study in New Zealand (T.E.M., associate director) started with a population-representative birth cohort and has sustained 94% participation for 5 decades by guaranteeing participants security of their confidential data through managed access and by investing in human techniques to build participant loyalty[53].

Discussion

Inequalities and discrimination remain pervasive in global society. Generational changes in disease prevalence and changes in the nature of disease conditions are also likely. Added to these facts are the changed economic circumstances now that global society is in the living-with-COVID-19 era. Given these challenges going forward, will LSVD be the best investment for human societies? Indeed, one such exercise has been abandoned (the Taiwan Biobank[54,55]). Is it time for researchers, funders and policy makers to step back and think about value for money in LSVD investments and the future costs of sustaining LSVD infrastructures? We think so. What types of research do societies need to address the United Nations' Sustainable Development Goals in relation to population health and well being? The urgency does not relate only to the continued investment in databanks. The highly standardized and technologically oriented data collection of LSVD perhaps leads to a loss of 'heart and soul', dislocated from human lives as humans live them. Data-mining experts who are not involved in the data-collection process are unlikely to have knowledge of the individuals contributing data or their communities, cultures and contexts[56]. Many

LSVD users are lacking in experience of what it is to collect primary data from people in place, and they often publish findings as given facts without adequate interpretations of their meaning for societies. There is a place for LSVD, although so many, perhaps not? We argue for a clear-eyed view of the relative value and contributions of different types of research to society, now and into the future, to assess the balance of investment and needs of future populations.

The COVID-19 pandemic has highlighted the vital importance of public trust in science and highlighted that science must be relevant for the populations most at risk. Unfortunately, healthy volunteer and other types of recruitment biases inherent to many LVSD may lead to the public's perception that medical research, and therefore public research investment, is only for the benefit of relatively privileged groups in society. Such a perception could further erode the public's trust in science. In this time of heightened public sensitivity to structural racism and social exclusion, LSVD could also become a red flag for social movements' concern about research as these movements aim to improve diversity and inclusion. LSVD's lack of representation of ethnic groups and unhealthy vulnerable groups of people is relevant, as is LSVD's lack of investigation of the quality and meaning of lives lived in their social, cultural and community contexts. And, although truly new findings have emerged from these datasets, arguably these do not touch the irrefutable worldwide evidence that inequalities and disadvantages are the primary drivers for poor health and mental well being for everyone and for unhealthy aging in whole communities[57]. Increasing numbers of sweeps, intensity of investigations and follow-ups in unrepresentative LSVD cannot redress inherent biases and omissions of factors that cause disease in people vulnerable to unhealthy aging. Our view is that associations derived from LSVD cannot guide clinical medicine and public health, as their outputs imply and as audiences infer, for all age, gender, ethnic and socioeconomic groups.

A further concern noted above regarding LSVD, given the consuming investment into such studies, is the assumption that disease and disease causation remain the same across time, in different generations and in different population groupings. Diseases and their risk factors can differ markedly across generations, but there has been little attention to how so-called historical cohort effects might impact underlying biology. For example, there have been historical changes in the occurrence of some cancers (deaths from those not related to exposure to tobacco are rising), asthma (rising in midlife), chronic obstructive pulmonary disease, stroke (reduction in ischemic varieties), heart attacks (change in electrocardiogram phenotype), diabetes (rising type II diabetes) and dementia (reduction in incidence)[56,58,59]. It seems entirely reasonable to hypothesize that, just as infectious diseases change across time in their relation to human hosts, so do chronic diseases and their biologies. The starkest example at present is the evolution of infectious organisms in relation to how we interact with other species and our environment, such as *Escherichia coli* pathogenicity and antimicrobial resistance. As well as these historical shifts, routine administrative records very clearly reflect clinical fashions and changing practice, such as changes in diagnostic criteria (for example, refs. [18,60]), often driven by the commercial sector and vested interests. So, just because a label appears in a databank's medical records over the years, it cannot be assumed to be one 'gold standard' that defines the same disorder across places, generations or time. It seems that much basic epidemiological training has been abandoned in this regard and needs to be re-emphasized. LSVD should prepare themselves and their users to study historical shifts in diseases, as well as their causation, given the clear changes in the population occurrence of many chronic diseases and the opportunity that this poses to discover new causes.

For those databanks that are currently funded and ongoing with their massive and dedicated teams and an understandable voracious need for continued funding, what is possible? Undoubtedly knowledge is emerging from databanks that is perceived to be of value. We point out here that researchers and science consumers need to understand that exact value better. Acknowledging limitations forthrightly in all publications from databanks is an important step (Table 1) but just a first step. There should be a deeper reflection, questioning assumptions about the wider contribution of these LSVD to human societies. This is a research question in itself. The answers would allow outputs to be contextualized and interpreted appropriately, whether it be by politicians, funders, journal editors or researchers. It is important that societies retain diversity in the nature of their research, and societies must invest in this diversity, not just LSVD monocultures. It is also important that societies retain pipelines of future researchers across disciplines who pay close attention to challenges to human health and well being, within and across social groupings. Early-career researchers should have boots-on-the-ground experiences of measuring disease and aging, not just workstation experience.

Research needs to shift toward informing primary prevention at scale in communities, which is where people age and become unwell. Research that is valuable for aging societies would address inequalities and social drivers of poor health, changes across generations and differences across communities, cultures and ethnic groups, factors that are currently poorly measured by routine medical records and data-capture technologies that are disconnected from peoples' lives. Topics of valuable research will include the geriatric giants such as falls, incontinence, confusion, dementia, frailty and subjective quality of life. Measuring these topics is quite different from measuring biomedical and clinical metrics, although such metrics can be integrated to allow a richer understanding, from genotype and phenotype to community, locality and population.

More broadly, databank investors and leaders need to consider how health research funding can contribute to the United Nations' Sustainable Development Goal missions, as well as the now fully recognized climate crisis. This essay provides a challenge to databank leaders and journal editors who publish databank findings to create more transparency and clarity about LSVD and their limitations. Future outputs from LSVD research are likely to be orientated to personalized 'early detection' of diseases and disease treatments, enterprises that are eagerly anticipated to benefit the pharmaceutical and technological industries[61], and therefore to benefit national treasuries through revenues. However, it is important to reflect on the future that this implies, a highly biomedicalized, technologized and quantified human life course sustained at considerable cost of personalized disease treatments at the individual level. This scenario does not align well with the already compelling evidence that recommends primary disease prevention at the population level. Personalized medicine will also likely increase cost to the environment, because it is already known that our health care systems, whether in diagnostics or treatments, contribute substantially to increasing global heating as well as depletion of natural resources and energy[12]. We call on funders to consider an appropriately framed review of LSVD that covers contributions made to date, including truly new findings versus repeating already established findings, cost up to the present and possible scenarios for the future including environmental costs of the research itself and the societal and environmental costs of the anticipated types of outcome, including more medicalization of our future selves by diagnostic and treatment technologies[62–64].

Such a future does not align with global societal challenges including addressing climate change, environmental degradation and global inequalities. This Perspective invites researchers and funders to reflect on the kind of evidence base needed for future population health and well being in the context of the Sustainable Development Goals[42]. We should get beyond the shared delusion that a big *N* alone will guarantee the accuracy of findings, their

relevance for vulnerable groups and their future usefulness for society. It is not enough.

Received: 30 July 2021; Accepted: 2 August 2022;
Published online: 13 September 2022

References

1. Chalmers, D. et al. Has the biobank bubble burst? Withstanding the challenges for sustainable biobanking in the digital era. *BMC Med. Ethics* **17**, 39 (2016).
2. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
3. Chen, Z. et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
4. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
5. Tyrrell, J. et al. Genetic predictors of participation in optional components of UK Biobank. *Nat. Commun.* **9**, 886 (2021).
6. Pirastu, N. et al. Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* **53**, 663–671 (2021).
7. All of Us Research Program Investigators et al. The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
8. Lynn, P. & Borkowska, M. *Some Indicators of Sample Representativeness and Attrition Bias for BHPS and Understanding Society*. Report No. 2018-01 (University of Essex, 2018).
9. Taylor, A. E. et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **47**, 1207–1216 (2018).
10. Cornish, R. P., Macleod, J., Boyd, A. & Tilling, K. Factors associated with participation over time in the Avon Longitudinal Study of Parents and Children: a study using linked education and primary care data. *Int. J. Epidemiol.* **50**, 293–302 (2020).
11. Jockwitz, C., Wiersch, L., Stumme, J. & Caspers, S. Cognitive profiles in older males and females. *Sci. Rep.* **11**, 6524 (2021).
12. *Cancer Disparities* (National Cancer Institute, 2022) https://www.cancer.gov/about-cancer/understanding/disparities
13. Dornquast, C. et al. Strategies to enhance retention in a cohort study among adults of Turkish descent living in Berlin. *J. Immigr. Minor. Health* **24**, 1309–1317 (2022).
14. Moffitt, T. E., Belsky, D. W., Danese, A., Poulton, R. & Caspi, A. The Longitudinal Study of Aging in Human Young Adults: knowledge gaps and research agenda. *J. Gerontol. A Biol. Sci. Med. Sci.* **72**, 210–215 (2017).
15. Baldwin, J. R., Reuben, A., Newbury, J. B. & Danese, A. Agreement between prospective and retrospective measures of childhood maltreatment: a systematic review and meta-analysis. *JAMA Psychiatry* **76**, 584–593 (2019).
16. Ramirez, A. H., Gebo, K. A. & Harris, P. A. Progress with the All of Us Research Program: opening access for researchers. *JAMA* **325**, 2441–2442 (2021).
17. Matthews, F. E. et al. A two-decade comparison of prevalence of dementia in individuals aged 65 years and older from three geographical areas of England: results of the Cognitive Function and Ageing Study I and II. *Lancet* **382**, 1405–1412 (2013).
18. Brayne, C., Gao, L., Dewey, M. & Matthews, F. E. Dementia before death in ageing societies—the promise of prevention and the reality. *PLoS Med.* **3**, e397 (2006).
19. Shiekh, S. I. et al. Ethnic differences in dementia risk: a systematic review and meta-analysis. *J. Alzheimers Dis.* **80**, 337–355 (2021).
20. Matthews, F. E. et al. A two decade dementia incidence comparison from the Cognitive Function and Ageing Studies I and II. *Nat. Commun.* **7**, 11398 (2016).
21. Matthews, F. E. et al. Epidemiological pathology of dementia: attributable-risks at death in the Medical Research Council Cognitive Function and Ageing Study. *PLoS Med.* **6**, e1000180 (2009).
22. Schneider, J. A., Arvanitakis, Z., Bang, W. & Bennett, D. A. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology* **69**, 2197–2204 (2007).
23. Fisch, L. et al. Predicting brain-age from raw T_1-weighted magnetic resonance imaging data using 3D convolutional neural networks. Preprint at https://arxiv.org/pdf/2103.11695 (2021).
24. Adams, M. J. et al. Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *Int. J. Epidemiol.* **49**, 410–421 (2020).
25. Davis, K. A. S. et al. Mental health in UK Biobank – development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. *BJPsych Open* **6**, e18 (2020).
26. Barr, P. B., Bigdeli, T. B. & Meyers, J. L. Prevalence, comorbidity, and sociodemographic correlates of psychiatric disorders reported in the All of Us Research Program. *JAMA Psychiatry* **79**, 622–628 (2022).
27. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.* **1**, 123–126 (2012).
28. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* **368**, m131 (2020).
29. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
30. Pizzi, C. et al. Sample selection and validity of exposure–disease association estimates in cohort studies. *J. Epidemiol. Community Health* **65**, 407–411 (2011).
31. Richiardi, L., Pizzi, C. & Pearce, N. Commentary: representativeness is usually not necessary and often should be avoided. *Int. J. Epidemiol.* **42**, 1018–1022 (2013).
32. Rothman, K. J., Gallacher, J. E. J. & Hatch, E. E. Why representativeness should be avoided. *Int. J. Epidemiol.* **42**, 1012–1014 (2013).
33. Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
34. Keyes, K. M. & Westreich, D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* **393**, 1297 (2019).
35. Mayeda, E. R. et al. Can survival bias explain the age attenuation of racial inequalities in stroke incidence? A simulation study. *Epidemiology* **29**, 525–532 (2018).
36. Swanson, J. M. The UK Biobank and selection bias. *Lancet* **380**, 110 (2012).
37. *What is MELODEM?* (BU Epidemiology, accessed 1 September 2022); http://sites.bu.edu/melodem/what-is-melodem/
38. Holmberg, M. J. & Andersen, L. W. Collider bias. *JAMA* **327**, 1282–1283 (2022).
39. Griffith, G. J. et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* **11**, 5749 (2020).
40. Kaplan, R. M., Chambers, D. A. & Glasgow, R. E. Big data and large sample size: a cautionary note on the potential for bias. *Clin. Transl. Sci.* **7**, 342–346 (2014).
41. Faber, J. & Fonseca, L. M. How sample size influences research outcomes. *Dental Press J. Orthod.* **19**, 27–29 (2014).
42. *Researcher Workbench* (All of Us Research Hub, 2022); https://www.researchallofus.org/data-tools/workbench/
43. Eisner, M. & Fearon, P. Pitfalls of using data portals as sources for psychological research: the example of cross-national homicide data. *Psychological Sci.* **32**, 863–865 (2021).
44. Simmons, C. et al. Responsible use of open-access developmental data: the Adolescent Brain Cognitive Development (ABCD) Study. *Psychol. Sci.* **32**, 866–870 (2021).
45. Kuss, O. et al. Statistical analysis in the German National Cohort (NAKO) – specific aspects and general recommendations. *Eur. J. Epidemiol.* **37**, 429–436 (2022).
46. Bennett, D. A. et al. Development of the standards of reporting of neurological disorders (STROND) checklist: a guideline for the reporting of incidence and prevalence studies in neuroepidemiology. *Eur. J. Epidemiol.* **30**, 569–576 (2015).
47. Tchetgen Tchetgen, E. J. Identification and estimation of survivor average causal effects. *Stat. Med.* **33**, 3601–3628 (2014).
48. Szklo, M. Population-based cohort studies. *Epidemiol. Rev.* **20**, 81–90 (1998).
49. Kearney, P. M. et al. Cohort profile: the Irish Longitudinal Study on Ageing. *Int. J. Epidemiol.* **40**, 877–884 (2011).
50. Norberg, S. J., Toohey, A. M., Jones, S., McDonough, R. & Hogan, D. B. Examining the municipal-level representativeness of the Canadian Longitudinal Study on Aging (CLSA) cohort: an analysis using Calgary participant baseline data. *Health Promot. Chronic Dis. Prev. Can.* **41**, 48–56 (2021).
51. Poulton, R., Moffitt, T. E. & Silva, P. A. The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Soc. Psychiatry Psychiatr. Epidemiol.* **50**, 679–693 (2015).
52. Libby, P. The changing landscape of atherosclerosis. *Nature* **592**, 524–533 (2021).
53. Marmot, M. Health equity in England: the Marmot review 10 years on. *BMJ* **368**, m693 (2020).
54. Corlateanu, A. et al. 'Chronic obstructive pulmonary disease and phenotypes: a state-of-the-art'. *Pulmonology* **26**, 95–100 (2020).

55. Harmon, S. H. E., Yen, S.-Y. & Tang, S.-M. Biobank governance: the cautionary tale of Taiwan Biobank. *SCRIPTed* https://doi.org/10.2966/scrip.150118.103 (2022).
56. *Asthma Statistics* (British Lung Foundation, 2022); https://statistics.blf.org.uk/asthma
57. Aldus, C. F. et al. Undiagnosed dementia in primary care: a record linkage study. *NIHR Journals Library* https://doi.org/10.3310/hsdr08200 (2020).
58. *Biobanking: Technologies and Global Markets* (BCC Research, 2016); https://www.bccresearch.com/market-research/biotechnology/biobanking-technologies-markets-report.html
59. *Sustainable Development Unit Archive* (NHS, 2009); https://www.england.nhs.uk/greenernhs/whats-already-happening/sustainable-development-unit-archive/
60. *17 Goals to Transform Our World. Sustainable Development Goals* (United Nations, 2016); https://www.un.org/sustainabledevelopment/
61. *Oxford English Dictionary* (Oxford University Press, 2022).
62. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
63. Metten, M.-A., Costet, N., Multigner, L., Viel, J.-F. & Chauvet, G. Inverse probability weighting to handle attrition in cohort studies: some guidance and a call for caution. *BMC Med. Res. Methodol.* **22**, 45 (2022).
64. Lin, J.-C., Fan, C.-T., Liao, C.-C. & Chen, Y.-S. Taiwan Biobank: making cross-database convergence possible in the big data era. *GigaScience* **7**, gix110 (2018).

Acknowledgements

T.E.M. is supported by grants AG032282, AG073207 and AG069939 from the National Institute on Aging and grant MR/P005918/1 from the UK Medical Research Council. C.B.'s cohort work has been supported by grants from multiple funders including the MRC, ESRC, NIHR, Alzheimer's Society and ARUK.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Carol Brayne.

Peer review information *Nature Aging* thanks Daniel Belsky and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.