International Journal of Epidemiology, 2018, 654–668
doi: 10.1093/ije/dyx283
Advance Access Publication Date: 8 February 2018
Education Corner

Education Corner

Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities

Catherine R Lesko,[1]* Lisa P Jacobson,[1] Keri N Althoff,[1] Alison G Abraham,[1,2] Stephen J Gange,[1] Richard D Moore,[1,3] Sharada Modur[1] and Bryan Lau[1]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, [2]Division of Ophthalmology and [3]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA

*Corresponding author. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., E7139, Baltimore, MD 21205, USA. E-mail: clesko2@jhu.edu

Editorial decision 12 December 2017; Accepted 3 January 2018

Abstract

Collaborative study designs (CSDs) that combine individual-level data from multiple independent contributing studies (ICSs) are becoming much more common due to their many advantages: increased statistical power through large sample sizes; increased ability to investigate effect heterogeneity due to diversity of participants; cost-efficiency through capitalizing on existing data; and ability to foster cooperative research and training of junior investigators. CSDs also present surmountable political, logistical and methodological challenges. Data harmonization may result in a reduced set of common data elements, but opportunities exist to leverage heterogeneous data across ICSs to investigate measurement error and residual confounding. Combining data from different study designs is an art, which motivates methods development. Diverse study samples, both across and within ICSs, prompt questions about the generalizability of results from CSDs. However, CSDs present unique opportunities to describe population health across person, place and time in a consistent fashion, and to explicitly generalize results to target populations of public health interest. Additional analytic challenges exist when analysing CSD data, because mechanisms by which systematic biases (e.g. information bias, confounding bias) arise may vary across ICSs, but multidisciplinary research teams are ready to tackle these challenges. CSDs are a powerful tool that, when properly harnessed, permits research that was not previously possible.

Key words: Cohort studies, collaborative study design, data harmonization, heterogeneity, pooled analyses

Key Messages

- Collaborative study designs (CSDs), in which independent contributing studies (ICSs) linked by a scientific commonality, perhaps with heterogeneous study designs and data-collection protocols, agree to share data, can become fertile environments for scientific productivity.
- Data harmonization across CSDs is often both retrospective and prospective; combining these approaches may leverage information across ICSs to quantify, and perhaps correct, measurement error and residual confounding.
- It is not a simple task to identify the target population to whom results from an analysis nested in a CSD can trivially generalize. However, the increased participant diversity in a CSD may allow investigators to describe heterogeneity in outcomes of interest across person, place and time; may lead to increased precision in estimating subgroup effects of treatments/exposures; and may permit results to be quantitatively generalized to key target populations of public health interest.
- Analytic approaches for CSDs will vary based on the research question and logistics, such as privacy concerns, data availability and characteristics of the ICSs included in the study sample.
- In addition to acknowledged strengths, CSDs present political, logistical and methodological challenges that are serious, but not insurmountable.

Introduction

The term 'collaborative study design' (CSD) may encompass a wide variety of models, from a multi-site randomized trial with standardized protocols, to a network of observational studies, perhaps with heterogeneous designs. Herein, we focus on collaborations between multiple independent contributing studies (ICSs), established with the intent of addressing multiple scientific questions, where research protocols are not standardized across ICSs, but harmonization of core data elements is feasible. CSDs present many strengths and opportunities, as well as surmountable methodological, political and logistical challenges. As CSDs become more common, identifying and addressing common challenges may help guide future collaborative initiatives. Herein, we: (i) discuss methodological challenges and opportunities when analysing data from multiple ICSs, (ii) review the state of the science for addressing these challenges and (iii) very briefly describe the political and logistical considerations when establishing the scaffolding for collaborative research.

Defining CSD

A sufficient set of conditions for a CSD includes: (i) a scientific commonality that unites the various studies, which may be a common population (e.g. children[1]), common outcome (e.g. HIV[2] or childhood cancer[3]) or a common exposure (e.g. genetics[4]), (ii) overlapping data elements, termed the 'core' data elements for the CSD and (iii) buy-in from leaders of the ICSs. Typically, ICSs agree to share individual-level data, which are then harmonized and pooled into a single dataset. However, CSDs may also undertake 'collective analyses' in which individual sites follow the same analysis plan and then site-specific results are meta-analysed. We distinguish this collective approach from a traditional meta-analysis in that: included studies are those that agree to collaborate, rather than those who have previously published results relevant to the question under investigation; the development of the analysis plan is generally preceded by harmonizing needed variables including outcome, exposures and covariates; and all sites agree to execute the same analysis plan (limiting heterogeneity due to different covariate adjustment sets or inclusion/exclusion criteria).[5] Some examples of existing CSDs of ICSs appear in Table 1. We also provide links to public websites, which include examples of governance processes for successful collaboration, and published research that illustrates many of the analytic challenges and opportunities outlined below.

Factors contributing to the proliferation of CSDs

Over the last two decades, several factors have contributed to the advent of the 'mega-cohort'[6,7] and the explosion of CSDs for epidemiologic research. Increased scientific interest in precision (or personalized) medicine, subgroup-specific effects, very broad arrays of risk factors such as genome-wide association studies and other '-omics', and clinically important rare exposures (e.g. multidrug resistant HIV) and rare events (e.g. HIV treatment failure)[8] have necessitated larger study samples than are available in typical, single-site epidemiological studies. Furthermore, funding agencies have also promoted collaboration in an effort

Table 1. Examples of collaborative studies

Subject/collaboration	Year established	Number of studies[a]	Number of individuals[a]	Collaborative model	Considerations	Website
Cardiovascular						
Asia Pacific Cohort Studies Collaboration (APCSC)[74]	~1999	44	600 000+	Data are not available for use by researchers outside of the APCSC, and researchers within the APCSC wishing to work on the data have to do so in Sydney, with prior approval from the Executive Committee	Restricted to large cohort studies; broad scope of outcomes, yet not all cohorts collect all necessary data and no standardization for what is collected	http://www.apcsc.info/
Cancer						
Vitamin D Pooling Project (VDPP)[75]	2007	10	4539 cases	A subset of cohorts from the National Cancer Institute Cohort Consortium; nested case–control study using a central laboratory and standardized testing protocols	Formed with a very specific goal: to address the gap in knowledge about association between vitamin D and rarer cancer sites	https://epi.grants.cancer.gov/VitaminD/
Collaborative Group on Hormonal Factors in Breast Cancer[42]	~1996	~54	147 000+	All studies that included ≥100 women with breast cancer with information on the use of hormonal contraceptives and reproductive history	Cohort studies were included using a nested case–control design, matching four random controls to each case	NA
Breast and Prostate Cancer Cohort Consortium (BPC3)[76]	2003	10	30 000+	A subset of cohorts from the National Cancer Institute Cohort Consortium; case–control study where cases came from member cohorts and consortium funding helped enrol additional cases/controls	Case–control study to identify genetic variants associated with cancer risk	https://epi.grants.cancer.gov/BPC3/
Pooling Project of Prospective Studies of Diet and Cancer[77]	1991	16	1 026 547	Summary data from prospective cohorts with validation data on dietary assessment data analysed centrally using two-stage analysis (random effect for cohort), standardized for all analyses	Comprehensive analysis plan published in *American Journal of Epidemiology* (2006)	https://www.hsph.harvard.edu/pooling-project/
Child health						
Environmental influences on Child Health Outcomes (ECHO)[47]	2016	84 cohorts funded in 35 awards by NIH	~50 000	Diverse set of existing cohorts will harmonize extant data and agree to prospective collection of new data elements; participating cohorts have access to extensive laboratory support and expertise in measuring	ECHO collaboration also includes Institutional Development Award (IDeA) States Pediatric Clinical Trials Network, ideally to be able to test interventions that are hypothesized to improve child	http://www.echochildren.org/

(continued)

Table 1. Continued

Subject/collaboration	Year established	Number of studies[a]	Number of individuals[a]	Collaborative model	Considerations	Website
				participants reported outcomes; research primarily structured around key outcomes—pre-, peri- and postnatal outcomes, airway conditions and sleep, neurodevelopment, obesity and positive health	health based on observational study findings in medically underserved and rural populations	
Diabetes						
EURODIAB ACE[78]	1993	~24	~30 million	Arose out of a prior collaboration established in 1985 (EURODIAB); collaborations monitored population incidence rates of childhood diabetes across Europe	Main outcome was rate of diabetes, which was susceptible to differential ascertainment; all study sites applied capture-recapture methods to monitor and report ascertainment completeness	
Genetics						
Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium[4]	2008	5(+)	38 000	Series of prospectively planned joint meta-analyses of GWAS; studies opt in to analyses and cohorts beyond the five official members are often invited as co-collaborators (co-collaborators often collaborated with CHARGE cohorts prior to the formation of CHARGE)	Studies used different genotyping platforms that resulted in <60 000 SNPs in common; need for phenotype standardization across cohorts; scope of collaboration exceeded scope of individual studies and stretches limited resources (e.g. need to collect phenotype data beyond what was originally proposed); within-study analysis followed by between-study meta-analysis avoids need for individual-level data sharing	http://www.chargeconsortium.com/
HIV						
Antiretroviral Therapy Collaborative Cohort (ART-CC)[79]	2000	19	12 574–74 000+	Cohorts submit data according to a protocol that standardizes variable definitions, formats and coding; all analyses are approved by a steering committee	Lots of heterogeneity in patient characteristics (% female, % IDU, AIDS, CD4+ cell count at baseline) and attrition rates (2–18%)	http://www.bristol.ac.uk/art-cc/
Concerted Action on SeroConversion on AIDS and Death in Europe	1997	28	25 000+	A cohort of seroconverters (subgroups of existing cohorts); anonymized data	Identification of seroconverters and estimated date of seroconversion varies across studies; collaborative	http://www.ctu.mrc.ac.uk/our_research/research_areas/hiv/studies/cascade/

(continued)

Table 1. Continued

Subject/collaboration	Year established	Number of studies[a]	Number of individuals[a]	Collaborative model	Considerations	Website
(CASCADE) Collaboration[80]				are collected; all analyses are approved by a steering committee	is part of EuroCoord, a 'Network of Excellence' that provides scientific oversight for CASCADE and other large cohorts and collaborations (www.eurocoord.net)	
Centers for AIDS Research Network of Integrated Clinical Systems (CNICS)[38]	2006	8	30 000+	Data extraction of a pre-defined set of variables; anonymized prior to submission to Data Coordinating Site; all analyses are approved by a steering committee; individual cohorts opt in to participate in any proposed analysis	Includes historical data on patients in care from 1995+ for some sites; collaboration also coordinates research on stored specimen maintained in repositories at each individual study site	https://www.uab.edu/cnics/
Collaboration of Observational HIV Epidemiological Research Europe (COHERE)[81]	2005	40	300 000+	Uses HIV Cohorts Data Exchange Protocol[82]—a standardized method of data structure and transfer—to compile data	Includes other, smaller collaboratives within it—e.g. CASCADE, EuroSIDA, UK CHIC	http://www.cohere.org/
HIV Cohorts Analyzed Using Structural Approaches to Longitudinal Data (HIV-CAUSAL)[83]	~2010	12	63 000+	Originally formed to answer: When to start ART? What regime to start? And when to switch regimes?	Explicitly created to be analysed using causal methods; includes historical data on patients in care from 1996 to 1998+ for all cohorts	https://www.hsph.harvard.edu/miguel-hernan/hiv-causal-collaboration/
International Epidemiologic Databases to Evaluate AIDS (IeDEA)[2,84,85]	2005	>50	1 700 000+	This is a collaboration of regional collaborations across the world, encompassing North America, Central and South America, Asia and Pacific Islands, and three African regions;[7] analyses must be approved by regional steering committees or, to use data from entire collaborative, by an executive committee	Scale of heterogeneity increased due to increased number of sites; includes other collaboratives (sometimes collaboratives of collaboratives) within it—e.g. NA-ACCORD is the North American regional representative	http://www.iedea.org/
Kidney disease						
Chronic Kidney Disease Prognosis Consortium[86]	2009	46	2 100 000+	Cohorts can share either individual de-identified data or can run a standard programme to create all output needed to include their study in meta-analytic results	Restricted to prospective studies with ≥1000 participants (except CKD cohorts) and ≥1 outcome of interest with a minimum of 50 events	http://www.jhsph.edu/research/centers-and-institutes/chronic-kidney-disease-prognosis-consortium/

[a]Most recent or approximate estimate; most cohorts are open and continue to enrol members and most collaboratives continue to enrol (and lose) individual studies.
CKD, Chronic Kidney Disease; GWAS, Genome-Wide Association Studies.

to maximize return on existing research infrastructure and data.[9] Finally, increased computing power and cheaper electronic storage have lowered technological barriers to curating and analysing large datasets.

The factors driving the proliferation of CSDs reflect their obvious advantages. Foremost is the large combined cohort size and the associated increased precision of estimates that address research questions that could not be answered in a single study. Additionally, increased diversity of participants in CSDs allows more extensive descriptive analyses by person, place and time, as well as the investigation of potentially meaningful effect heterogeneity. This strength is especially evident in the case when subgroups at highest risk for the outcome of interest are traditionally underrepresented in ICSs. Geographical diversity allows for mapping of indicators. Trends over time are distinguishable when ICSs with overlapping but extending time frames are united. Some of the most informative papers from CSDs carefully describe key disease indicators across important subgroups, a wide geographical-spread or over the course of time.[10,11] Collaboration is also cost-efficient, maximizing the usefulness of existing data by encouraging secondary data analysis. Finally, collaborations have the potential to become fertile environments for scientific productivity and for training of new investigators due to the availability and accessibility of secondary data,[12] coupled with scientific guidance and input from senior experts in the field, who often serve as principal investigators of ICSs.

Methodological considerations for CSDs

CSDs often form to address multiple research questions related to a common theme, yet each specific question addressed in a CSD may require a slightly different methodological approach. Aspects of study design that may vary with the structure of the CSD and the research question include: defining the target population and inclusion/exclusion criteria for creating the study sample; defining key covariates, including the exposure and/or outcome of interest; and determining an analysis plan. Some steps of this process—e.g. defining key covariates—may be standardized across analyses nested in the CSD as part of the data-harmonization process.

Target populations and study samples

Generalizability

Generalizability is not a feature of a particular study sample, but rather describes a relationship between a study sample and a well-defined target population.[13–15] There may be a different target population for every unique research question nested in a CSD. In theory, to answer a research question, we would enrol a probability sample from the target population of interest to ensure generalizability of results to that target population. In practice, CSDs leverage data from existing independent studies that may or may not have study samples drawn from the target population that is truly of interest, and then determine to whom results can be generalized. Because defining the study sample is a multi-step, complicated process (described below), additional effort, often beyond that required for a single study, is required to posit generalizability of results to a particular target population. Despite this challenge, the heterogeneity of participants in CSDs beyond that usually seen in single studies often means the study sample in a CSD may be more representative of key target populations of interest[10] or that results can be formally reweighted to target populations of interest to present results explicitly generalized to that target population[16] under a well-defined set of sufficient assumptions and with information on the distribution of key covariates in the target population.[14,15] Whereas explicitly addressing the generalizability of results to a particular target population arguably increases their utility, there may be instances in which representativeness and generalizability are not of primary concern, e.g. when first evaluating a hypothesis, exploring heterogeneity or when specific interest is in subgroups.[17,18]

Selection processes leading to the study sample

Study sample: describing the participant pool in a CSD

The following selection processes and criteria determine the participant pool in a CSD from which more restricted study samples can be chosen to answer a specific nested research question: (i) the ICSs that were *invited* or *eligible* to join the CSD, (ii) the eligible ICSs that *chose* or were *selected* to participate in the CSD, (iii) ICSs' original sampling mechanisms (informed by inclusion/exclusion criteria and the source populations from which participants were recruited) and (iv) additional selection criteria applied to ICS study samples when determining eligibility for inclusion in the CSD (obtaining additional participant consent for data sharing, requiring participants be alive and still under follow-up, etc.) (Figure 1).

To obtain the most diverse study sample possible in a CSD, and thus increase the chances of being able to generalize to a particular target population (see below), ideally, a CSD would invite all studies that share the scientific commonality the CSD is attempting to leverage and that have collected the core data elements of interest.[19] However, there may be political, logistical or resource limitations to this ideal, resulting in only a subset of studies being

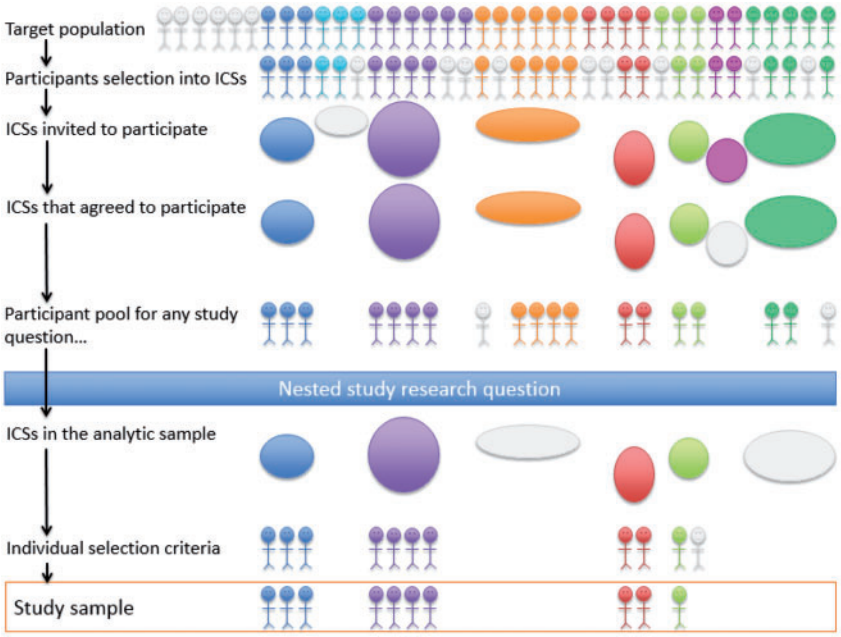

Figure 1. Relationship between target population and study sample for answering a research question nested in a collaborative study design.

extended an invitation.[2] Some CSDs have enrolled only 'large' studies,[19] presumably because data harmonization is time- and resource-intensive and coordination across studies can be complex. For example, ICSs were chosen to participate in the Environmental influences on Child Health Outcomes (ECHO) cohort study (http://echochildren.org), through a competitive process run by the National Institutes of Health (https://www.nih.gov/echo). By design, recruitment was limited to ICSs whose study participants resided in the USA and its territories. Interested ICSs were encouraged to assemble cohorts, either currently active or inactive but where participants may be contacted for re-enrolment, in their applications to yield scientifically justifiable large sample sizes and be willing to prospectively collect data elements consistent with the aims of the ECHO's scientific agenda.

In a CSD in which all ICSs were extant prior to formation of the CSD, ICSs may have employed unique sampling mechanisms and sampled from varied source populations. Sometimes, these sampling mechanisms may be similar enough to be modelled in pooled data to assess generalizability to a particular target population.[16] To our knowledge, methods and implications for generalizing results from a CSD of ICSs with substantially different sampling mechanisms have not been explored. Stratified sampling models are likely necessary, if a reweighting approach is to be undertaken (as has been described for generalizing results from a single study sample). If study participants in the CSD cannot be thought of as a biased sample of the target population, methods for transportability (rather than generalizability) should be considered.[15,20,21] If some ICSs sampled from source populations nested within the target population and some did not, the simplest solution may be to restrict the primary analysis to those ICSs that sampled from the target population and explore the sensitivity of results to participant characteristics, locale and health systems by conducting separate analyses in those ICSs that did not.

Study sample: identifying eligible ICSs for a specific research question

After some sets of ICSs are invited to and agree to participate in the CSD, participating ICSs may be given the opportunity to opt in or out of specific CSD nested studies. The study sample for a CSD nested study must be understood as a compilation of study samples of the participating ICSs, which are defined by their individual source populations and sampling mechanisms. It is easy to give the false impression that a study nested in a CSD is based upon a clearly defined study sample and target population. Pooling of data across diverse subgroups may add to the strength of the CSD, especially if heterogeneity of results is explicitly investigated, but investigators must also be aware of any imbalances in the analytic sample due to the over-influence of an ICS or subgroup. Subgroup-specific analyses should be routinely undertaken and reported, either as a primary analysis, as an interim step in the analytic process or as part of a sensitivity analysis.

Study sample: identifying eligible participants for a specific research question

The final step in defining the study sample for a CSD nested study is applying individual-level inclusion/exclusionary criteria to participants in ICSs that have opted into

the study. Inclusion/exclusion criteria may apply to participant demographics, clinical characteristics or time. For example, if the nested study is longitudinal, study entry and exit dates are identified for each individual according to an algorithm that incorporates: the study period or originating or landmark event specific to the research question, enrolment date (which itself may be determined by an algorithm based on enrolment into an ICS, date the ICS began participating in the CSD and individuals' study visit attendance), death date, lost-to-follow-up date and outcome date.

Considerations in selecting the study sample may vary based on the target population and estimand of interest, as identified by the research question. For example, if the research question is descriptive or predictive, the goal may be to obtain a sample in which the research question may be addressed while minimizing selection bias and maintaining generalizability (applicability) to some target population. If the research question asks whether a causal relationship exists between some exposure and some outcome, additional consideration must be paid to confounding control. These objectives may sometimes be at odds with one another. A common strategy for selecting the study sample is to restrict to individuals with non-missing data for key variables; this *may* result in greater internal validity (under restrictive assumptions about the missing data mechanism)[22,23] but less power and less external validity; alternatively, this strategy may induce selection bias.[24,25] Assembling the study sample to answer a research question using a CSD requires subject-matter knowledge of the variables affecting the outcome of interest, the exposure(s) (if applicable) and selection.[14,25–27]

Data harmonization

Data harmonization may be retrospective (for extant data) or prospective (for data collected after the CSD is formed) and stringent or flexible.[28] Most CSDs are based on existing studies with established protocols and extant data, and thus most data harmonization is retrospective, although, after the formation of a CSD, ICSs may agree to also harmonize data collection prospectively. By necessity, because most data harmonization is retrospective, it is also most commonly flexible. The degree of flexibility may have inferential implications for studies nested in CSDs. Data harmonization is complex and a full treatment of the process and tools[29,30] is beyond the scope of this paper; instead, we highlight several implications of data-harmonization decisions on analyses conducted in CSDs.

Typically, aspects of extant core data elements will vary across ICSs. Variables may be recorded differently, measured differently or may measure slightly different constructs.[31] For example, one ICS may measure smoking history as ever/never at baseline, collected via audio computer-assisted self-interview, whereas another may measure pack-years of smoking, collected via face-to-face interview. Data harmonization of extant data across ICSs often defaults to the lowest common denominator for a variable (in the example above, collapsing pack-years of smoking at baseline into ever/never). However, this approach may result in loss of information. For variables that are deemed confounders for a particular study question, this approach may result in residual confounding. If instead, the original variables were modelled within each ICS and results meta-analysed, the degree of confounding control would differ across cohorts and the importance of tight confounding control may be evident in heterogeneity in site-specific estimates of effect.[32] Sensitivity analyses, validation studies and methods for correcting for measurement error that results from the data-harmonization process are appropriate in these settings.[33–35] A third option is to treat the more detailed data (e.g. pack-years of smoking) as missing for studies that collected less detailed data (e.g. ever/never smoking) and use other studies with complete data as validation subsamples for correcting for missing data and measurement error.[34,36,37] A fourth option is to restrict the study sample to ICSs with comparable data elements to answer the research question at hand. This is common practice for CSDs: almost every research question relies on a subset of ICSs that have similar data on exposure, outcome and key covariates of interest. The first three approaches to data harmonization are more often considered for covariates that are of secondary interest, although their use is not restricted to these settings. If data harmonization is undertaken at the inception of a CSD, meta-data can be helpful for investigators not involved in the data-harmonization process to understand the data elements available to them in the harmonized dataset.

It is not uncommon for CSDs to undertake collection of new data elements to augment existing data as important gaps in the common dataset are realized. Typically, the protocol for new data collection is developed through a cooperative process that balances individual ICS resources and objectives with common CSD goals. For example, the Center for AIDS Research Network of Integrated Clinical Systems (CNICS) includes eight HIV clinical cohorts that initially relied on retrospective data harmonization to merge common data elements related to HIV medical care (e.g. laboratory data, antiretroviral therapy use data, demographics).[38] Since its establishment, many CNICS sites have started collecting patient-reported outcomes (PROs)[39,40] guided by a commonly-agreed-upon set of data elements; however, each ICS retains ultimate discretion over the actual questions asked and assessment tools

used at their site. For questions that are not of primary interest to the CSD, the PRO platform allows ICSs to pursue their own research agendas. Frequently, questions piloted by individual sites end up as part of the set of common data elements. More stringent approaches to prospective data collection are possible.

Harmonizing extant data and implementation of new prospective data-collection protocols approaches have their unique strengths and limitations.[28] The retrospective data harmonization may be a strength of CSDs in that more detailed information from some ICSs may be leveraged to complement data from ICSs that collect less detailed information and variations in results due to different data-collection methods can be explored. One strength of prospective data collection is that newly collected data elements could be leveraged to better quantify relationships between different data-collection mechanisms by serving as an internal validation study, if e.g. the two different measurements were collected on the same sample. A challenge to prospective data collection is the resource requirements both of ICS investigators and participants. Furthermore, if data elements change over time, assessing temporal trends or isolating associations with those variables over time may be more challenging. The combination of retrospective and prospective data collection presents an opportunity for methods development to ensure the accuracy of studies nested in CSDs.[41]

Combining data from different study designs

CSDs often combine data from different study designs, including clinical and interval cohorts,[2] or from cohort and case–control studies.[42] There are underlying characteristics of the study samples captured by such study designs that may be important, unmeasured sources of participant heterogeneity.[43,44] Interval and population-based cohorts typically include volunteers on whom standardized data are collected at regular intervals and for whom retention is typically high due to purposeful follow-up and participant tracking.[44] In contrast, clinical or administrative cohorts may include individuals with greater need for and access to care, on whom data are not standardized or collected at regular intervals and for whom retention is typically lower.[44] Finally, case–control data are often collected retrospectively based on self-report or administrative records, and rarely include a longitudinal component unless the study is nested within a cohort study. Depending on how cases and controls are ascertained, case–control studies may come from a less clearly defined target population, which could be a source of bias, particularly if cases come from a tertiary care setting.[45,46] However, for rare outcomes or exposures that are expensive or invasive to collect, case–control studies are efficient.[47,48]

The study design of the ICSs and its relationship to the research question may influence whether data from an ICS can be used in a study nested in the CSD. Part of the art of a CSD is combining data in smart and thoughtful ways. When the research question at hand calls for a cross-sectional analysis, this could be as simple as combining cross-sectional study data with cohort study data, reduced to the cross-sectional subset at baseline or other meaningful time point. When the research question requires a longitudinal analysis, more sophisticated approaches are possible, such as combining cross-sectional data on current duration from the time origin to the event of interest and cohort data on time to the event, using a current duration analysis[49] and survival analysis, respectively. If the research question can be answered with a case–control study or if the majority of ICSs that might be able to contribute used a case–control design, extant case–control studies may be combined with cohort studies by nesting a case–control study within the cohort study.[42,50,51] The drive to include as many ICSs as possible in a given analysis motivates much methods development in CSDs.

Analytical considerations

After data have been harmonized, a research question clearly articulated and a study sample clearly defined, there are sometimes multiple options for how to analyse data in a CSD (i.e. across ICSs). In a 'collective' or 'disseminated analysis', data are analysed within each ICS and then summary results are analysed using meta-analysis techniques. Alternatively, in a 'pooled analysis', individual-level data from each ICS may be combined into a single dataset and then analysed as if it were one cohort (accounting for within-ICS correlation). Technological advances have also made it possible to conduct pooled analyses without physically transferring and pooling data.[52] Importantly, from a statistical standpoint, all else being equal, collective analyses and pooled analyses produce the same or similar results.[53,54] Thus, typically, logistics, including the ease or state of data sharing and concerns about data privacy, security and ownership,[55] may drive investigators to analyse data using a collective, rather than a pooled, approach. Alternatively, some research questions may not lend themselves to collective analyses and a pooled analysis will be preferred or required, e.g. when the outcome is so rare that individual ICSs will not have enough events to conduct independent analyses for meta-analysis or they will have too few events to support confounder control.

Collective analyses

In collective analyses, site-specific analyses are combined using a meta-analytic approach. Site-specific analyses may be

conducted centrally or locally at each site ('disseminated analysis'). In a disseminated approach, the primary investigator or data-analysis centre typically supplies code to each site to ensure analyses are uniform across sites. One possible challenge for disseminated analyses is that all analyses must be agreed upon a priori and exploratory analyses are logistically infeasible; in addition to summary results, ICSs should also submit interim descriptive statistics so that any potential problems combining final results can be identified. If site-specific analyses are to be conducted centrally, individual-level data are required. Once data-sharing arrangements are in place, most commonly this individual-level data will include all variables necessary for analysis (since often the CSD will have created a central, harmonized dataset from which analytic datasets may be generated for research questions as they arise). However, if participant privacy is a concern, sites can share individual-level data on only outcome, exposure and propensity scores, with all other identifying information stripped from the data. A propensity score is an individual probability of exposure, conditional on a set of covariates thought sufficient to control confounding. Adjusting for, standardizing on, stratifying on or matching on propensity score are all methods for confounder control. The benefits of sharing propensity scores include: if any confounding variables are personally identifying, they do not need to be shared; the data-analysis centre can explore the impact of different analytic decisions (e.g. different calipers for matching) without requiring additional effort from ICSs; and propensity scores can be estimated locally at each site so that sites retain ownership of their data.[55–58] If ICSs submit individual-level data with propensity scores, analyses should still be stratified by ICS because a propensity score at one site may not represent the same propensity at another site.[55,59]

Pooled analyses of individual-level data

Pooled, individual-level data may tempt investigators into treating data as if they arose from a single study; however, significant heterogeneity may still exist across ICSs. Furthermore, ICSs are typically based in distinct geographic regions or clinics, or share some other eligibility criteria; thus, participants within an ICS will be more similar to one another than they will be to participants in another ICS. Analysing data as if coming from one study sample (i.e. ignoring the differences in source populations and study designs) can lead to single summary effect estimates that are misleading (if there is important heterogeneity by ICS), biased (if the ICS is associated with both exposure and outcome) or overly precise (if observations are correlated within ICS and the analysis does not account for that correlation).[60]

Confounding

Pooled analyses are often restricted to set of covariates common to all sites. In contrast, analyses stratified by site (whether conducted centrally or using a disseminated approach) might control for a standard set of key covariates available across all sites or the full set of measured confounders available at each site.[55,57] Adjusting for all confounders available at each site may reduce residual confounding, but may decrease interpretability of results (because the adjustment set varies across the study). This approach also reduces feasibility of sensitivity analyses for unmeasured confounding (because they would need to be conducted separately within each ICS depending on the ICS-specific adjustment set). However, it is possible that fully adjusted estimates could be leveraged to further adjust estimates from other sites that did not measure all desired confounders.[61] The choice of adjustment set did not make a substantive difference in results in one example[62] but, to our knowledge, different scenarios have not been tested with a simulation, and the optimal choice may vary with the situation.

If heterogeneity across ICSs is not of interest, e.g. when the factors that may be driving differences can be explicitly studied or if they are so nebulous as to be uninterpretable, ICS may confound the analysis and should be in the adjustment. Table 2 demonstrates how confounding by ICS can be present, even when ICS-specific effect estimates are unbiased. In Table 2, within each ICS, the distribution of causal types[63] (descriptors of individuals' potential outcomes under both exposure levels) is balanced between exposed and unexposed persons, i.e. exposed and unexposed groups are exchangeable. Yet, because the prevalence of exposure varies across ICS (67% in Study A and 33% in Study B), as does the risk of the outcome (15% among the unexposed in Study A vs 30% in Study B), naively pooling individuals from the two studies results in confounding. This confounding is evident in the imbalance of causal types in the combined sample, as well as the non-collapsibility of the risk difference and risk ratio (also the odds ratio, but the odds ratio may not be collapsible even in the absence of confounding).[64] A common source of confounding by ICS is differences in the observation window covered by each ICS. If observation windows across ICSs differ across calendar time, there may be important cohort effects that could be controlled for by controlling for ICS. If observation windows across ICSs differ by disease stage or some other biologically relevant time-scale, there may be important frailty effects (unmeasured prognostic indicators that may also influence treatment propensity)[65] that could be controlled for by controlling for ICS. Observation windows can be established by reviewing protocols of the ICSs and using data-visualization

Table 2. Example of how ICS may confound pooled analyses in the absence of confounding within ICS; table entries are *N* (column per cent) unless otherwise noted

	Study A		Study B		Combined sample	
	Exposed	Unexposed	Exposed	Unexposed	Exposed	Unexposed
Causal response type[a]	2000 (100)	1000 (100)	1000 (100)	2000 (100)	3000 (100)	3000 (100)
Doomed	200 (10)	100 (10)	300 (30)	600 (30)	500 (17)	700 (23)
Causal	1000 (50)	500 (50)	450 (45)	900 (45)	1450 (48)	1400 (47)
Protective	100 (5)	50 (5)	0 (0)	0 (0)	100 (3)	50 (2)
Immune	700 (35)	350 (35)	250 (25)	500 (25)	950 (32)	850 (28)
Risk[b]	60%	15%	75%	30%	65%	25%
Risk difference	45%		45%		40%	
Risk ratio	4.0		2.5		2.6	
Odds ratio	8.5		7.0		5.6	

[a]Causal response types summarizes individuals' potential outcomes had they been exposed and had they been unexposed. Doomed individuals would have the outcome regardless of exposure status; immune individuals would not have the outcome regardless of exposure status; causal individuals would only experience the outcome if they were exposed; and protective individuals would only experience the outcome if they were unexposed.

[b]Risk is determined based on combinations of the causal response types and actual exposure status. In the subset of the sample that is exposed, individuals who are doomed and causal would be observed to experience the outcome. In the unexposed subset, individuals who are doomed and protective would experience the outcome.

techniques such as Lexis diagrams.[66] Increases, decreases and plateaus of effects found in analyses may be artifacts of ICSs entering and exiting at times other than the study period start and stop dates.

The following analytical considerations generally apply to both collective and pooled, individual-level analyses.

Measurement error

The accuracy of variable measurement may vary across ICSs. For example, some ICSs may link to registries to capture events (e.g. deaths, cancers) in participants who are lost to follow-up, whereas others may not. Furthermore, even if death ascertainment were complete across studies, some studies may record cause of death from death certificates whereas other studies may adjudicate cause of death using death certificates, medical records and interviews with proxies; sensitivity and specificity of cause-specific mortality vary with both data source and cause.[67] If measurement of a key variable is hypothesized to be a source of heterogeneity or bias, it may be reasonable to restrict analyses to data from ICSs that follow a particular protocol for measuring that variable. However, this approach has the potential to restrict sample size. The bias reduction that would result must be balanced against the reduction in power that would come from excluding cases.[68]

Different frequency of variable capture means investigators must choose between coarsening the data of the cohorts with more frequent ascertainment or extrapolating data from the cohorts with less frequent ascertainment. Either strategy may introduce measurement error. The degree of measurement error may depend on how variable or stable the covariate is over time.[43] When outcomes are not detectable outside of a clinical encounter (e.g. HIV RNA viral rebound, onset of diabetes), their documented incidence or prevalence will vary according to study participants' visit frequency and differences in study protocol for detection. For example, if a particular condition was not the outcome of interest in the original ICS, it may not have been ascertained at every visit. In clinical cohorts particularly, the visit structure may be informative.[69] That is, sicker patients may be more likely to have outcomes detected because they are more likely to seek care or get tested. There are several analysis options to control for variable observation patterns, including joint longitudinal/survival analysis[70] or inverse probability of observation weighting.[69,71]

Missing data

Variable values may be missing by design (e.g. because a protocol did not require their collection) or they may be inadvertently missing (e.g. because a laboratory specimen was lost or a participant declined to answer a question). In a complete cohort analysis, researchers must choose between restricting analysis to a sample on whom all variables are collected (and potentially introducing selection bias and reducing power) or neglecting to control for some confounders (and potentially being left with residual confounding bias). If methods (e.g. multiple imputation or inverse probability weighting) are used to handle missing data, the predictors of missingness[23] may differ by study. Even if data are missing at random across all studies, the variables that inform the missingness may be different and

attempts to correct for missingness may need to be study-specific. If multiple imputation is used,[72] more research is needed to provide guidance as to when to pool across all ICSs before imputing, when ICS should be included as a predictive variable in the imputation or when imputation should be completely stratified by ICS. Stratifying by ICS would allow the missing-data mechanism to vary across cohorts, and subsequent stratified analysis and meta-analytic pooling of cohort-specific results. In contrast, if the missing-data mechanism is the same across cohorts, pooling the data may maximally leverage all available information and decrease the variance.

Political and logistical challenges

Complete details on the political and logistical challenges of CSDs are beyond the scope of this paper, but we would be remiss if we did not acknowledge the great care that must be taken to establish a collaborative culture through the governance structure, internal processes and scientific partnerships of a CSD.

In order to be scientifically productive, CSDs require dedication and motivation from the investigators leading the ICSs, which must be earned by convincing them that collaboration adds value to and does not threaten the objectives of the individual ICSs. Common value-added benefits of collaboration may include: (i) a forum to discuss gaps in scientific knowledge and often additional administrative or technical resources to advance research to address those gaps, (ii) resources to overcome possible limitations of individual studies in the CSD (e.g. a lack of generalizability, an insufficient number of events, etc.) and/or a larger platform to validate and extend the ICS's research and (iii) a rich environment for training junior investigators that provides access to data and a network of content-specific experts.

ICSs may hesitate to collaborate if they feel that collaboration will reduce their agency in deciding how their data are used, or threaten their brand or their ability to achieve their own scientific aims. This concern can be mitigated through a strong, fair and explicit governance structure, explicit policies for use of data and a formal research approval process. Governance through a scientific steering committee assists these efforts, particularly when the committee includes at least one representative from each ICS. When the CSD has been convened by an agency such as the National Institutes of Health, including an agency partner on the scientific steering committee can be an asset by providing 'big picture' vision and direction. Furthermore, processes to gain general approval, use of data, engage in the design, analysis and interpretation of results, and review scientific products from the CSD, as well as criteria for authorship on CSD products, should be agreed to in writing at the onset of the collaboration. Particularly when collaborations are multinational, the processes governing data sharing in a CSD may be restricted by countries' privacy laws. Authorship would ideally be decided based on International Committee of Medical Journal Editors' (or other specific journal) guidelines;[73] other members of the CSD (e.g. principal investigators of ICSs) whose contributions do not justify authorship may agree to be listed in the 'Acknowledgements' section or as a member of the collaboration. Finally, the most productive culture in CSDs is one of inclusion. Ensuring processes that welcome investigators of the ICSs to contribute not only data, but scientific knowledge and opinion, at *all* steps in the research, is key to scientifically productive CSDs.

In addition to a culture of inclusivity, the other necessary (and primary) component for successful and sustained collaboration is the promise of opportunities to 'do good science'. The scientific questions asked in CSDs should add important information above and beyond what any one ICS could contribute. The value of descriptive studies in CSDs should not be underestimated, particularly when there is diversity in the CSD data by person, place and time. The possibility to engage in rigorous science that expands upon the work possible in the ICSs is one of the most appealing benefits of collaboration to participating investigators. To summarize, these and other political challenges of CSDs are best overcome with a culture of inclusivity born from a solid governance structure.

Very briefly, logistical processes that must be addressed for successful collaboration (if applicable) include but are not limited to: managing Institutional Review Board approval across ICSs and the CSD; summarizing specifics of available data within each ICSs (often referred to study-level 'meta-data'); managing a biospecimen repository (physical or virtual) and determining who should have priority for using limited stored samples; prioritizing approved studies for analyses; ensuring nested studies follow best practices for reproducible research; managing communication across the CSD components that include the ICSs; and managing the brand identity of the CSD while highlighting the ICSs. These and other logistic challenges of CSDs are overcome with strong, collaborative and transparent administration.

Conclusions

Aggregating data from multiple cohort studies provides a powerful tool that, when properly harnessed, allows research that was not previously possible. Methodological advantages to CSDs include increased power for rare outcomes, rare exposures and subgroup analyses; more informative epidemiologic descriptions by person, place and

time; potential for more diverse study samples and detection of biologically relevant heterogeneity; and the ability to maximize the potential of data that have already been collected. Designing nested studies in CSDs has added layers of complexity, and handling systematic biases resulting from measurement error, missing data and residual confounding also becomes more complicated, because the mechanisms by which they arise and their relative degree may vary across the study. Fortunately, multidisciplinary researchers are primed to tackle these challenges and embark on cutting-edge science. Whereas existing CSDs have done pioneering work to harmonize data and foster collaborative scientific productivity, the complexities and challenges of CSDs can spur innovation that leads to development of new and novel methods. The challenges of CSDs should be embraced as another opportunity for advancing science.

Funding

This work was supported by funding from the National Institutes of Health: U01 HL121812, U01 AA020793, UM11 AI035043, R24 AI067039, U01 AI042590, U01 AI069918 and U24 OD023382.

Conflict of Interest: Keri Althoff has served on scientific advisory boards for TrioHealth and Gilead Sciences, Inc. No other authors have conflicts of interest to declare.

References

1. Currie C, Nic Gabhainn S, Godeau E, International HNCC. The health behaviour in school-aged children: WHO Collaborative Cross-National (HBSC) study: origins, concept, history and development 1982–2008. *Int J Public Health* 2009;**54(Suppl 2)**: 131–39.
2. Gange SJ, Kitahata MM, Saag MS *et al*. Cohort profile: the North American AIDS Cohort Collaboration on Research and Design (NA-ACCORD). *Int J Epidemiol* 2007;**36**:294–301.
3. Brown RC, Dwyer T, Kasten C *et al*. Cohort profile: the International Childhood Cancer Cohort Consortium (I4C). *Int J Epidemiol* 2007;**36**:724–30.
4. Psaty BM, O'Donnell CJ, Gudnason V *et al*. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation Cardiovascular Genetics* 2009;**2**:73–80.
5. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol* 1999;**28**:1–9.
6. Gaziano JM. The evolution of population science: advent of the mega cohort. *JAMA-J Am Med Assoc* 2010;**304**:2288–89.
7. Sorlie P, Wei GS. Population-based cohort studies: still relevant? *J Am Coll Cardiol* 2011;**58**:2010–13.
8. Deeks SG, Gange SJ, Kitahata MM *et al*. Trends in multidrug treatment failure and subsequent mortality among antiretroviral therapy-experienced patients with HIV infection in North America. *Clin Infect Dis* 2009;**49**:1582–90.
9. Roger VL, Boerwinkle E, Crapo JD *et al*. Strategic transformation of population studies: recommendations of the working group on epidemiology and population sciences from the National Heart, Lung, and Blood Advisory Council and Board of External Experts. *Am J Epidemiol* 2015;**181**:363–68.
10. Althoff KN, Buchacz K, Hall HI *et al*. U.S. trends in antiretroviral therapy use, HIV RNA plasma viral loads, and CD4 T-lymphocyte cell counts among HIV-infected persons, 2000 to 2008. *Ann Intern Med* 2012;**157**:325–35.
11. Levey AS, de Jong PE, Coresh J *et al*. The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. *Kidney Int* 2011;**80**:17–28.
12. Longo DL, Drazen JM. Data sharing. *N Engl J Med* 2016;**374**: 276–77.
13. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol* 2010;**172**:107–15.
14. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology* 2017;**28**:553–61.
15. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci USA* 2016;**113**:7345–52.
16. Lesko CR, Cole SR, Hall HI *et al*. The effect of antiretroviral therapy on all-cause mortality, generalized to persons diagnosed with HIV in the USA, 2009–11. *Int J Epidemiol* 2016;**45**:140–50.
17. Rothman K, Hatch E, Gallacher J. Representativeness is not helpful in studying heterogeneity of effects across subgroups. *Int J Epidemiol* 2014;**43**:633–34.
18. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;**42**:1012–14.
19. Collaborative Group on Epidemiological Studies of Ovarian C, Beral V, Doll R, Hermon C, Peto R, Reeves G. Ovarian cancer and oral contraceptives: collaborative reanalysis of data from 45 epidemiological studies including 23,257 women with ovarian cancer and 87,303 controls. *Lancet* 2008;**371**:303–14.
20. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *J Causal Inference* 2013;**1**: 107–34.
21. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol* 2017;**186**:1010–14.
22. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res* 2012;**21**:243–56.
23. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996;**91**:473–89.
24. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;**15**:615–25.
25. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology* 2012;**23**:159–64.
26. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;**155**:176–84.
27. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
28. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: consolidating data harmonization—how to obtain quality and applicability? *Am J Epidemiol* 2011;**174**:261–64; author reply 5–6.

29. Fortier I, Burton PR, Robson PJ *et al*. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39**:1383–93.
30. Granda P, Blasczyk E. Data harmonization. In: Center. UoMSR (ed). *Guidelines for Best Practice in Cross-Cultural Surveys*, 2nd edn. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan, 2010.
31. van der Steen JT, Kruse RL, Szafara KL *et al*. Benefits and pitfalls of pooling datasets from comparable observational studies: combining US and Dutch nursing home studies. *Palliat Med* 2008; **22**:750–59.
32. Sang Y, Matsushita K, Mahmoodi BK, Astor BC, Coresh J, Woodward M. Abstract P343: Comparison of two-stage and one-stage meta-analyses: an example of eGFR-Cardiovascular Mortality Association (for CKD-PC collaborators). *Circulation* 2012;**125**(**Suppl 10**):AP343-AP.
33. Lau B, Gange SJ. Methods for the analysis of continuous biomarker assay data with increased sensitivity. *Epidemiology* 2004;**15**:724–32.
34. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol* 2006;**35**:1074–81.
35. Cole SR, Jacobson LP, Tien PC, Kingsley L, Chmiel JS, Anastos K. Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *Am J Epidemiol* 2010;**171**:113–22.
36. Edwards JK, Cole SR, Westreich D *et al*. Multiple imputation to account for measurement error in marginal structural models. *Epidemiology* 2015;**26**:645–52.
37. Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Stat Med* 2008;**27**:5195–216.
38. Kitahata MM, Rodriguez B, Haubrich R *et al*. Cohort profile: the Centers for AIDS Research Network of Integrated Clinical Systems. *Int J Epidemiol* 2008;**37**:948–55.
39. Crane HM, Lober W, Webster E *et al*. Routine collection of patient-reported outcomes in an HIV clinic setting: the first 100 patients. *Curr HIV Res* 2007;**5**:109–18.
40. Broderick JE, DeWitt EM, Rothrock N, Crane PK, Forrest CB. Advances in patient-reported outcomes: the NIH PROMIS((R)) measures. *EGEMS (Wash DC)* 2013;**1**:1015.
41. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *Int J Epidemiol* 2015;**44**:1452–59.
42. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: further results. *Contraception* 1996;**54**(**Suppl 3**):1S–106S.
43. Lau B, Gange SJ, Kirk GD, Moore RD. Evaluation of human immunodeficiency virus biomarkers: inferences from interval and clinical cohort studies. *Epidemiology* 2009;**20**:664–72.
44. Lau B, Gange SJ, Moore RD. Interval and clinical cohort studies: epidemiological issues. *AIDS Res Hum Retroviruses* 2007;**23**: 769–76.
45. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics* 1946;**2**:47–53.
46. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. *Am J Epidemiol* 1992;**135**:1029–41.
47. Wacholder S. Design issues in case-control studies. *Stat Methods Med Res* 1995;**4**:293–309.
48. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. *Principles*. *Am J Epidemiol* 1992;**135**:1019–28.
49. Keiding N, Kvist K, Hartvig H, Tvede M, Juul S. Estimating time to pregnancy from current durations in a cross-sectional sample. *Biostatistics* 2002;**3**:565–78.
50. Borgan O, Samuelsen SO. A review of cohort sampling designs for Cox's regression model: potentials in epidemiology. *Norsk Epidemiologi* 2003;**13**:239–48.
51. Collaborative Group on Hormonal Factors in Breast C. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet* 1996;**347**:1713–27.
52. Gaye A, Marcon Y, Isaeva J *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014;**43**:1929–44.
53. Olkin I, Sampson A. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* 1998;**54**: 317–22.
54. Steinberg KK, Smith SJ, Stroup DF *et al*. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *Am J Epidemiol* 1997;**145**:917–25.
55. Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care* 2010; **48**(**Suppl 6**):S83–89.
56. Fireman B, Lee J, Lewis N, Bembom O, van der Laan M, Baxter R. Influenza vaccination and mortality: differentiating vaccine effects from bias. *Am J Epidemiol* 2009;**170**:650–56.
57. Toh S, Platt R. Is size the next big thing in epidemiology? *Epidemiology* 2013;**24**:349–51.
58. Rauh VA, Andrews HF, Garfinkel RS. The contribution of maternal age to racial disparities in birthweight: a multilevel perspective. *Am J Public Health* 2001;**91**:1815–24.
59. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med* 2013;**32**: 2837–49.
60. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. *Ann Intern Med* 2001;**135**:112–23.
61. Chatterjee N, Chen YH, Maas P, Carroll RJ. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J Am Stat Assoc* 2016;**111**:107–17.
62. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010;**19**:848–57.
63. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol* 1986;**15**:413–19.
64. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14**:29–46.
65. Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations

in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol* 2010;**172**:843–54.
66. Lund J. Sampling bias in population studies—how to use the Lexis diagram. *Scand J Stat* 2000;**27**:589–604.
67. Hernando V, Sobrino-Vegas P, Burriel MC *et al*. Differences in the causes of death of HIV-positive patients in a cohort study by data sources and coding algorithms. *AIDS* 2012;**26**: 1829–34.
68. Abraham AG, D'Souza G, Jing Y *et al*. Invasive cervical cancer risk among HIV-infected women: *a North American multicohort collaboration prospective study. J Acquir Immune Defic Syndr* 2013;**62**:405–13.
69. Hernán MA, McAdams M, McGrath N, Lanoy E, Costagliola D. Observation plans in longitudinal studies with time-varying treatments. *Stat Methods Med Res* 2009;**18**:27–52.
70. Liu L, Huang X, O'Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* 2008;**64**:950–58.
71. Nevo ON, Lesko CR, Colwell B, Ballard C, Cole SR, Mathews WC. Outcomes of pharmacist-assisted management of antiretroviral therapy in patients with HIV infection: a risk-adjusted analysis. *Am J Health Syst Pharm* 2015;**72**: 1463–70.
72. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;**19**:618–26.
73. Defining the Role of Authors and Contributors. http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html (12 January 2018, date last accessed).
74. Woodward M, Barzi F, Martiniuk A *et al*. Cohort profile: the Asia Pacific Cohort Studies Collaboration. *Int J Epidemiol* 2006; **35**:1412–16.
75. Helzlsouer KJ, Committee VS. Overview of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. *Am J Epidemiol* 2010;**172**:4–9.
76. Hunter DJ, Riboli E, Haiman CA *et al*. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat Rev Cancer* 2005;**5**:977–85.
77. Smith-Warner SA, Spiegelman D, Ritz J *et al*. Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *Am J Epidemiol* 2006; **163**:1053–64.
78. Green A, Gale EA, Patterson CC. Incidence of childhood-onset insulin-dependent diabetes mellitus: the EURODIAB ACE Study. *Lancet* 1992;**339**:905–09.
79. May MT, Ingle SM, Costagliola D *et al*. Cohort profile: Antiretroviral Therapy Cohort Collaboration (ART-CC). *Int J Epidemiol* 2014;**43**:691–702.
80. Collaborative Group on AIDS Incubation and HIV Survival including the CASCADE EU Concerted Action. Concerted Action on SeroConversion to AIDS and Death in Europe. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. *Lancet* 2000;**355**:1131–37.
81. Collaboration of Observational HIVERESG, Sabin CA, Smith CJ *et al*. Response to combination antiretroviral therapy: variation by age. *AIDS* 2008;**22**:1463–73.
82. Kjaer J, Ledergerber B. HIV cohort collaborations: proposal for harmonization of data exchange. *Antivir Ther* 2004;**9**:631–33.
83. Collaboration H-C, Ray M, Logan R *et al*. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *AIDS* 2010;**24**:123–37.
84. Egger M, Ekouevi DK, Williams C *et al*. Cohort Profile: the international epidemiological databases to evaluate AIDS (IeDEA) in sub-Saharan Africa. *Int J Epidemiol* 2012;**41**:1256–64.
85. McGowan CC, Cahn P, Gotuzzo E *et al*. Cohort profile: Caribbean, Central and South America Network for HIV research (CCASAnet) collaboration within the International Epidemiologic Databases to Evaluate AIDS (IeDEA) programme. *Int J Epidemiol* 2007;**36**:969–76.
86. Matsushita K, Ballew SH, Astor BC *et al*. Cohort profile: the chronic kidney disease prognosis consortium. *Int J Epidemiol* 2013;**42**:1660–68.