An Ontology-based Approach to Guide and Document Variable and Data Source Selection and Data Integration Process to Support Integrative Data Analysis in Cancer Outcomes Research

Hansi Zhang[1], Yi Guo[1], Jiang Bian[1*]

[1] Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA

Email: Hansi Zhang – hansi.zhang@ufl.edu; Yi Guo – yiguo@ufl.edu; Jiang Bian – bianjiang@ufl.edu

* Corresponding author: Jiang Bian, bianjiang@ufl.edu.

Jiang Bian

Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, 2197 Mowry Road, Suite 122, PO Box 100177, Gainesville, FL 32610-0177

Abstract

Background

To reduce cancer mortality and improve cancer outcomes, it is critical to understand the various cancer risk factors (RFs) across different domains (e.g., genetic, environmental, and behavioral risk factors) and levels (e.g., individual, interpersonal, and community levels). However, prior research on RFs of cancer outcomes, has primarily focused on individual level RFs due to the lack of integrated datasets that contain multi-level, multi-domain RFs. Further, the lack of a consensus and proper guidance on systematically identify RFs also increase the difficulty of RF selection from heterogenous data sources in a multi-level integrative data analysis (mIDA) study. More importantly, as mIDA studies require integrating heterogenous data sources, the data integration processes in the limited number of existing mIDA studies are inconsistently performed and poorly documented, and thus threatening transparency and reproducibility.

Methods

Informed by the National Institute on Minority Health and Health Disparities (NIMHD) research framework, we (1) reviewed existing reporting guidelines from the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network and (2) developed a theory-driven reporting guideline to guide the RF variable selection, data source selection, and data integration process. Then, we developed an ontology to standardize the documentation of the RF selection and data integration process in mIDA studies.

Results

We summarized the review results and created a reporting guideline—ATTEST—for reporting the variable selection and data source selection and integration process. We provided an ATTEST check list to help researchers to annotate and clearly document each step of their mIDA

studies to ensure the transparency and reproducibility. We used the ATTEST to report two mIDA case studies and further transformed annotation results into sematic triples, so that the relationships among variables, data sources and integration processes are explicitly standardized and modeled using the classes and properties from OD-ATTEST.

Conclusion

Our ontology-based reporting guideline solves some key challenges in current mIDA studies for cancer outcomes research, through providing (1) a theory-driven guidance for multi-level and multi-domain RF variable and data source selection; and (2) a standardized documentation of the data selection and integration processes powered by an ontology, thus a way to enable sharing of mIDA study reports among researchers.

Keywords: Ontology, Integrative data analysis, Cancer outcomes research, Reporting guideline

BACKGROUND

Cancer is a major disease burden worldwide [1]. As the 2nd leading cause of death in the United States (US), about 1 in 4 deaths is due to various types of cancer [2]. In 2019, an estimation of 1,762,450 new cancer cases diagnosed and 606,880 cancer deaths is reported by the American Cancer Society (ACS) in US [2]. The lifetime probabilities of being diagnosed with cancer are 39.3% and 37.4% for male and female, respectively [3]. However, the risk factors for these high cancer incidence and mortality rates are still not fully understood.

Nevertheless, to reduce cancer mortality rates and improve cancer outcomes (e.g., survival and prognosis), it is critical to understand the various risk factors of cancer. So far, evidence suggests that it is the interaction among many risk factors (RFs) together that affect the risk of cancer and cancer outcomes, rather than a single cause [4]. Further, the RFs involved are across different domains (e.g., genetic, environmental, and behavioral risk factors) and levels (e.g., individual level, interpersonal level, and community level). However, there is not yet an agreement among the cancer research community regarding how these multi-level cancer RFs interact with each other. To do so, the first and most crucial step is to gain a comprehensive view of potential multi-level RFs associated with various cancer outcomes such as the stage of diagnosis (the most important prognostic factor) and survival.

We surveyed existing research on RFs for late stage cancer diagnosis and poor survival, we found current studies about RFs for cancer outcomes are mostly from single-level analyses with mostly individual patient-level data. For instance, Andrew *et al.* assessed individual patient characteristics (e.g., age, gender, family history), and lifestyle factors (e.g., education, insurance

and socioeconomic status) to study their risks associated with colorectal cancer at late stage [5]. These individual-level RFs have also been reported for other major types of cancers such as breast and cervical cancers [6–9]. Further, prior studies studying cancer RFs often only analyzed data from a single source, such as SEER [10], SEER-Medicare[11], or a state or hospital cancer registry [12]. Among these cancer risk factor studies, the complex interplay between difference levels RFs are often ignored (e.g., county-level smoking rate vs. individual smoking behavior). These single-level RF analyses (1) lead to biased effect estimates of RFs due to potential confounding from omitted factors, (2) omit critical cross-level RF interactions, such as race by residence, that could inform multi-level intervention design.

Nowadays, advances in technology created new ways for us to determine and measure disease risk factors across different levels (e.g., from advancements in genome sequencing for genetic markers to better sensors for producing more accurate estimates of environmental pollutants). The availability of such abundant data online in electronic formats enables researchers to pool data on an unprecedented scale and offers a great opportunity to do a thorough examination of multi-level RFs in a multi-level integrative data analysis (mIDA) so that confounding effects and across-level interactions can be studied. However, researchers face significant barriers to do so, especially because there is a lack of consensus and proper guidance to help researchers systematically think and discovery these variables from heterogenous sources. In 2017, National Institute on Minority Health and Health Disparities (NIMHD) of the National Institute of Health (NIH) proposed a Research Framework [13], an extension to the well-known social ecological model [14], to help investigators systematically study health disparities. Recognized by the NIMHD Framework, individuals are embedded within the larger social system and constrained

by the physical environment they live in. Within this framework, cancer outcomes are influenced by RFs from different levels (i.e., individual, interpersonal, community, and societal) and multiple domains (i.e., biological, behavioral, physical/built environment, sociocultural environment, and healthcare system). In this work, we adopted the NIMHD framework as the guiding theory for risk factor discovery and data source selection.

Further, mIDA for cancer outcomes research requires the integration of data from multiple sources. However, data integration processes in the very limited number of existing mIDA studies [15, 16] are inconsistently performed and poorly documented, and thus threatening transparency and reproducibility [17, 18]. The data integration processes are often time summarized in one or two sentences without explicitly documentation of the steps. For example, Guo *et al*. explored the impact of the relationships among socioeconomic status, individual smoking status, and community-level smoking rate on pharyngeal cancer survival [16]. The multi-level risk factors above were obtained and integrated from three different data sources (i.e., Florida Cancer Data System [FCDS], U.S. Census, and Behavioral Risk Factor Surveillance System [BRFSS]) as mentioned in the abstract. However, for the rest of the paper, there is no description of how the individual-level records from FCDS are linked with county-level smoking rate from BRFSS and census tract-level poverty rate from U.S. Census. Even though the integration process might be as simple as integrating these multi-level variables through the geographic code (e.g., county code), it still needs to be standardized and explicitly documented to avoid ambiguity. For example, the paper discussed that "*regional smoking was measured as the average percentage of adult current smokers at the county level between 1996 and 2010*" and the readers might be able to make an educated guess that the regional smoking rates were more

likely to be generated using the BRFSS data rather than from the FCDS data; however, explicit documentation is needed as both BRFSS and FCDS data have individual smoking status. Keegan *et al.* explored and whether breast cancer survival patterns are influenced by factors such as nativity (individual level) and neighborhood socioeconomic status (community level). Similarly, they summarized integration process in one sentence by stating each patient was assigned a neighborhood socioeconomic status variable based the census block groups. However, the details such as variable names in each data sources, or whether the original geographic variables require pre-processing (e.g., derive census tract from zip codes) are not clearly documented [15]. The explicit documentation of these variable selection and data integration processes will help readers to better understand the study results, benefit other researchers who want to replicate the studies, but also more importantly, make it possible for machines to understand and replicate the steps (when these explicit documentations are encoded in a computable format such as with an ontology).

Further, even though these mIDA studies above did not emphasize the need for data integration or integrated datasets, the fact that they can only investigated a handful of variables at a time indicated the lack of but needed support on data integration. Even in studies on building frameworks or platforms to support or automate the data integration process (especially those related to creating integrated dataset to support cancer research), they often ignored the need for documenting the integration steps to guarantee the transparency and reproducibility of their approaches. For example, semantic data integration approach —connecting variables across different databases at the semantic level through mapping them to standardized concepts in a global schema (e.g., often time a global ontology) — has been proposed in data integration

studies in recent years to support generating integrated datasets for cancer research [19–21]. However, none of these studies mentioned the need for standardizing and documenting their integration steps, for example, most of them did not even discuss the rationale for selecting the specific data sources to integrate. Nevertheless, when reporting mIDA studies, it is critical to document the steps that were followed to select, integrate, and process the data so that others can repeat the same steps and reproduce the findings.

To address challenges above, in this paper, we first developed a reporting guideline to guide and document the RF variable selection, data source selection, and data integration process. The guideline is informed by (1) the NIMHD research framework that provides guidance and promotes structural thinking on identifying multi-level cancer RFs; and (2) reviewing existing reporting guidelines from the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [22]. Then, we proposed an ontology-based approach to annotate and document the RF selection and data integration process in mIDA studies based on the reporting guideline we developed. To do so, we developed the **O**ntology for the **D**ocumentation of v**A**riable selec**T**ion and da**T**a sourc**E** **S**election and in**T**egration proess (OD-ATTEST) so that the RF selection and data integration report can be (1) explicitly modeled with a shared, controlled vocabulary, (2) understandable to humans and computable to computers, and (3) adaptive to changes when the reporting process is refined.

In our prior work [23], we proposed a preliminary reporting guideline for RF variable and data source selection based on our own experience of pooling multi-level RFs from different data sources to support mIDAs of cancer survival [24, 25]. In this extended journal paper, we

significantly expanded our ontology-based reporting guideline—ATTEST (v**A**riable selec**T**ion and da**T**a sourc**E** **S**election and in**T**egration):

- We conducted a systematic search of existing reporting guidelines from the EQUATOR network to extract reporting elements relevant to variable selection and data integration.
- We updated our reporting guideline based on the result of the systematic review to include new items regarding data integration (e.g., data processing, data integration strategy, data validation, etc.) as well as variable and data source selection.
- We completed building the OD-ATTEST following the best practice in ontology development to provide a formal presentation for the reporting guideline with standardized and controlled vocabularies.
- We provided an ontology (OD-ATTEST) annotated report generated based on a prior mIDA study to represent the annotated items and their relationships in reporting guideline.

METHODS

Development of a reporting guideline for risk factor selection, data source selection, and data integration

To develop the reporting guideline, we started with summarizing our previous studies where we assessed the effect of data integration on predictive ability of cancer survival models [24] and created a semantic data integration framework to pool multi-level RFs from heterogenous data sources to support mIDA [25]. In the above studies, we went through the process of RF selection, data source selection, and data integration. To be able to ensure the reproducibility of these studies, a number of middle steps need to be documented as detailed in our previous paper

[23]. For example, both rural-urban commuting area (RUCA) codes [26] and the National Center for Health Statistics (NCHS) urban-rural classification scheme [27] are often used to represent an geographic area's rurality status. The difference between the two resides in the classification granularity, where RUCA focuses on classifying U.S. census tracts (i.e., tens levels from rural to metropolitan) while the NCHS urban-rural classification scheme focuses on classifying U.S. counties (i.e., a hierarchal definition with six levels). Thus, we need to clearly document which rural definition we used in the data analysis since different representations of the same variable (i.e., rurality in this case) have different impacts on model results, as shown in our prior work [24]. Further, before integration RFs from various data sources at different levels (e.g., census tract level vs. county level) and covered different time periods, we made an assumption assumed that area-level characteristics (e.g., social vulnerability index) derived from 2000 U.S. Census data were applicable across different time periods (as our individual level data from FCDS covered 1996 and 2010). Above experiences suggest that we must document these data integration nuances so that other researchers can repeat our data integration and data processing pipeline and reproduce the same results (e.g., integrated dataset). In sum, we summarized 3 key items that need to be documented: (1) RF selection (e.g., individual vs county-level variables), (2) data source selection (e.g., individual-level data from FCDS and contextual-level data from US Census), and (3) data integration and data preprocessing strategies.

Through discussions with expert biostatisticians, data analysts, and cancer outcomes researchers, we summarized the typical mIDA process and found there is little structured thinking when investigators selecting and identifying risk factors and their data sources. We thus propose to use the NIMHD research framework to provide a theory-driven guidance for multi-level and multi-

domain RF and data source selections. The NIMHD framework is originally designed to depicts a wide range of health determinants (i.e., RFs from different levels and domains) relevant to understanding and addressing minority health and health disparities. The goal of using the NIMHD framework is to help investigators to structurally and comprehensively think and identify relevant RFs and corresponding data sources in their IDA studies.

To build upon existing established reporting guidelines, we searched and identified relevant reporting guidelines from the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network—a comprehensive searchable database of guidelines for health research reporting. The EQUATOR network categorizes health researches into 13 study types (e.g., quantitative studies, experimental studies, and observational studies), where reporting guidelines for observational studies are most relevant to our mIDA use case. To further identify relevant reporting guidelines in EQUATOR, we developed a set of screening criteria to determine whether a reporting guideline in EQUATOR contains the information that can be used to improve our ATTEST reporting guideline as shown below:

- The reporting guideline is designed for secondary data analysis studies.
- The reporting guideline contains at least one of the following sections: data, outcomes (variables), and methods, as these sections will contain information related to variable selection, data source selection, and data integration methods.
- The reported data within the guideline must be health related.
- The use of the guideline (at least part of the guideline) can be extended to the cancer outcomes research, especially those related to variable selection, data source selection, and data integration.

We reviewed all reporting guidelines designed for observational studies and eliminated guidelines that do not involve the tasks of RF and data source selection and integration. We then identified all reporting guidelines that contain the following sections: data, outcomes (variables), and methods. For those that do not have sections clearly marked, we manually reviewed the entire reporting guideline to identify whether they discussed one of the three aspects. We then extracted reporting items in the selected reporting guidelines that are relevant to RF selection, data source selection, and data integration. Two reviewers (HZ and JB) independently extracted these reporting items of interest and resolved conflicts with a third reviewer (YG). We further analyzed these extracted reporting items and discussed with experts (i.e., biostatisticians, data analysts and cancer outcomes researchers) to summarize items needed in our reporting guideline, especially those related to the data integration process.

Construction of an ontology for the documentation of variable and data source selection and integration process (OD-ATTEST)

The ATTEST reporting guideline we developed is used to guide the variable and data source selection and integration process in cancer outcomes research. We propose to use an ontology-based approach to annotate and document the items in the reporting guideline. The goal of the OD-ATTEST ontology is to standardize the terminology used in documenting the selection and integration steps of RF variables and data sources to support mIDA.

The OD- ATTEST is developed using Protégé 5. We used Basic Formal Ontology (BFO) [28] as the upper-level ontology. We first adopted a top down approach to enumerate important entities (classes and relations) based on the reporting guideline we developed. Following the

best practice, we reviewed existing widely accepted ontologies using the National Center for Biomedical Ontology (NCBO) BioPortal [29] to find the entities can be reused in OD-ATTEST. Then, we started with the definitions of the most general concepts in the domain and subsequent specialization of the concepts to develop the class hierarchy. We also took a bottom-up process, where we started with the definitions of the most specific classes, and then subsequent grouped similar classes into more general concepts. We also examined how these reporting items are associated with each other (e.g., "*sample size*" is determined by "*primary outcome*") and determined what additional classes and relations were needed to fully represent these entities in OD-ATTEST.

An OD-ATTEST-annotated report generated based on a mIDA case study following the reporting guideline

To test the developed ATTEST reporting guideline and the OD-ATTEST ontology, we first created a ATTEST report based on our previous mIDA case study, where we explored the impact of the relationships among socioeconomic status, individual smoking status, and community-level smoking rate on pharyngeal cancer survival [16]. To annotate the ATTEST report using OD-ATTEST, we used the following annotation process: 1) identify information related to the reporting items in ATTEST through reviewing the original publication and supplementary materials; 2) annotate the information using the entities in OD-ATTEST; and 3) transform annotation results into semantic triples in Resource Description Framework (RDF) format using Turtle syntax [30].

RESULTS

The ATTEST reporting guideline for RF variable and data source selection and data integration

We extended our preliminary reporting guideline [23] through a review of existing relevant reporting guidelines published in the EQUATOR network. ***Figure 1*** shows our review process. We reviewed 94 reporting guidelines designed for from observational studies in the EQUATOR network. Out of the 94 reporting guidelines, 30 contain the required data, outcomes (variables), and method sections, which we retained for data extraction. In the data extraction step, for each reporting guideline, we extracted items relevant to RF and data source selection and integration, where the data and outcomes (variables) sections often contain information regarding how RF variables and data sources are selected, while the method section contains information about how data are processed and integrated.

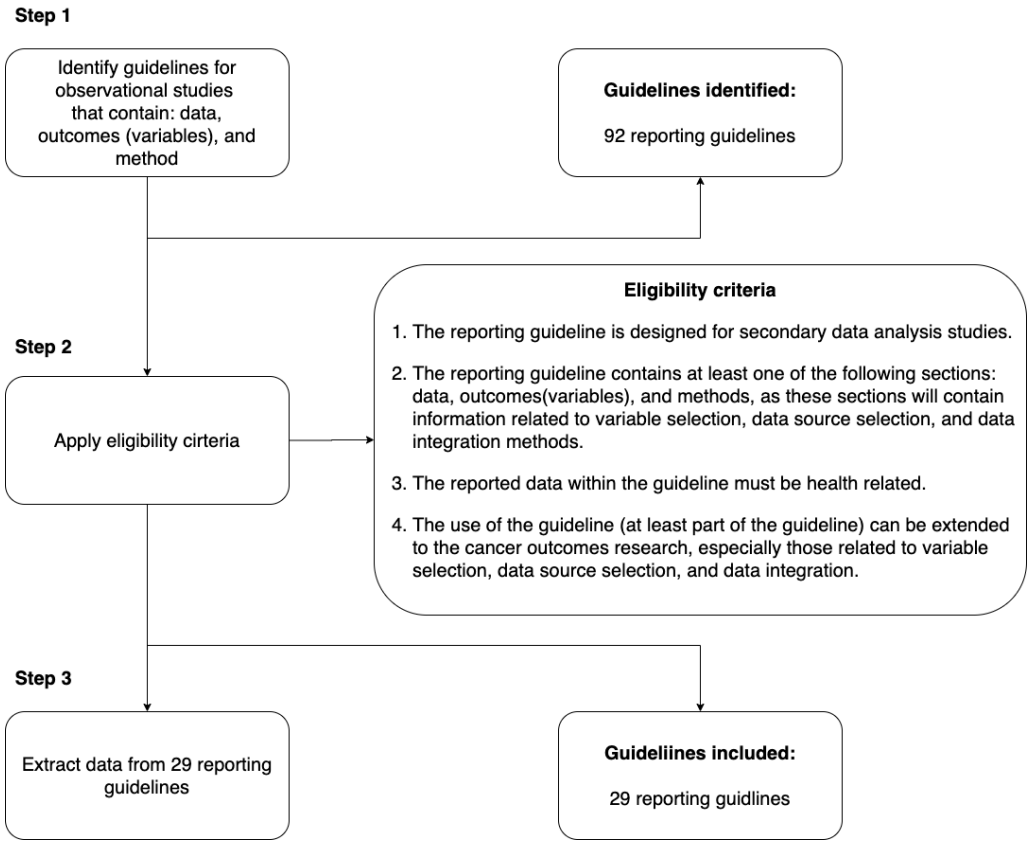

Figure 1. *Review of relevant reporting guidelines in the EQUATOR network.*

We categorized these reporting guidelines (***Table 1***) based on the domains and levels of the data sources reported in the guidelines and mapped them to the NIMHD framework. As shown in ***Table 1***, these 29 reporting guidelines cover data sources from all domains and levels of influences. Among them, 9 guidelines focused on providing a general reporting guideline for observational studies without specifying a specific domain of influence; while the rest of the guidelines are designed for different domains. For example, the Genetic RIsk Prediction Studies (GRIPS) statement [31] is designed for risk prediction studies using genetic data. Furthermore, most guidelines only considered the data sources from individual level, while 2 of them considered data sources from multi-levels. For example, the Checklist for One Health Epidemiological Reporting of Evidence (COHERE) [32] considered both individual and environmental risk factors when studying a disease.

Table 1. *Summary of reporting guidelines based on the data source domains and levels guided by the NIMHD framework.*

Domain of influences		**Level of influences**	**Guidelines**
Not specified*		Individual level	[33], [34], [35], [36], [37], [38], [39], [40], [41]
		Societal level	[42]
Biological data	Genetics data	Individual level	[31], [43]
	Immunogenomic data		[44]
	Molecular epidemiological data		[45], [46]
	Drug safety data from biologics registers		[47], [48]
Behavioral data	Crime, violence data	Individual level	[49]
	Dietary or nutritional data		[50]
	Medication adherence		[51]
Sociocultural environment	Environmental data	Individual/ Community/ Societal/Interpersonal	[32]
Physical environment			

Healthcare system	Administrative data, Electronic health records, Claim data, Patient or disease registries, Quality or safety surveillance databases	Individual level	[52], [53], [54], [55], [56] ,[57], [58], [59]
*When reporting data sources or RF variables, these studies did not specify a specific data domain.			

In our preliminary reporting guideline [23], we focused only on reporting items relevant to RF variables and data sources selection. In this review, we extracted items that can be used to improve our initial reporting guideline but with a focus on documenting the data integration process. In total, we found 3 reporting guidelines [53–55] contain information about data integration processes. However, items included in these 3 guidelines focus on data linkage and do not contain enough details about how to solve the heterogeneities of data from different sources. For example, when integrating variables across different levels (e.g., combine individual-level patient data and county-level smoking rate), none of the 3 guidelines have items on documenting the cross-level integration choices (e.g., layering the county-level smoking rate to individual based on residence of the individuals and county code), while this type of choices is frequently encountered in mIDA studies. Further, data processing steps such as the choices and algorithms used for creating new data elements (e.g., compute a body mass index variable from two separate variables, weight and height) are not documented in existing reporting guidelines. Therefore, we further extended the ATTEST to include these important data integration and data processing procedures based on our previous research experience on building data integration framework [25].

Informed by the NIMHD research framework and consistent with our prior work, the ATTEST reporting guideline consists of two main parts as shown in ***Figure 2***, reporting (1) the objective

of the study including explaining the background and rationale for designing the study in one or two sentences and describing the hypothesis of the study; and (2) the study design for variable and data source selection processes and describing the data along with the data integration and processing strategies. The variable and data source selection process consist of five key steps: (1) define the outcome variables for primary and (if necessary) secondary outcomes; (2) for each outcome variable, follow an iterative process (see ***Figure 2.A***) to determine the data sources according to NIMHD framework. After selecting each outcome variable and data sources, investigators need to think about how to select or consolidate similar outcome variables from the different selected data sources. For example, if the outcome of interest is an individual's lung cancer risk, we shall first identify potential data sources (e.g., cancer registries or electronic health records [EHRs]) that contain individual-level patient data where lung cancer incidence data are available. Then, based on the cohort criteria and other information such as required sample size and data range (e.g., time coverage and geographic information) of the potential data sources, the investigator could determine the qualified data sources and choose an adequate one based on the objective and design of the study. For example, if 2 data sources, cancer registry and EHRs, are both available and contain individual-level lung cancer incidence data, the investigator has the choices to (1) choose one data source over the other, or (2) link the two data sources and integrate variables from the two data sources. If the investigator chooses to link and integrate the two data sources, she needs to explicitly document the linkage and integration processes for each of variables as shown in ***Figure 2*** (Report – Variables – E, F, G, H) so that others can repeat the processes to generate the same analytical dataset; (3) determine the individual-level predictors and covariates of the study; (4) for each individual-level predictor or covariate, follow loop B in ***Figure 2*** to identify the different levels/domains of predictors or

368 covariates according to NIMHD framework. Similar to the outcome variables, different data
369 sources could potentially contain the same predictor or covariate variable, thus, it is important to
370 contrast and consolidate a new predictor or covariate with the existing selected predictors and
371 covariates to resolve duplicates. If an investigator chooses to integrate the "duplicate" variables
372 (e.g., choosing smoking status from cancer registry data over EHRs because cancer registries
373 data are manually abstracted and typically have better data quality than raw EHRs), these data
374 integration choices also need to be explicitly documented. Nevertheless, it is often a difficult
375 choice and these "duplicate" variables might all need to be tested in models before a selection
376 can be made. Regardless, these decisions and data processing steps need to be clearly
377 documented; and (5) after selecting individual-level predictors and covariates, one can use a
378 similar process, following loop C in ***Figure 2*** to identify additional contextual-level predictors
379 and covariates and data sources of interest. In the end, a report of the selected data and data
380 sources as well as the data integration processes shall be generated as shown in ***Figure 2***. The
381 corresponding ATTEST reporting guideline checklist is shown in ***Table 2.***

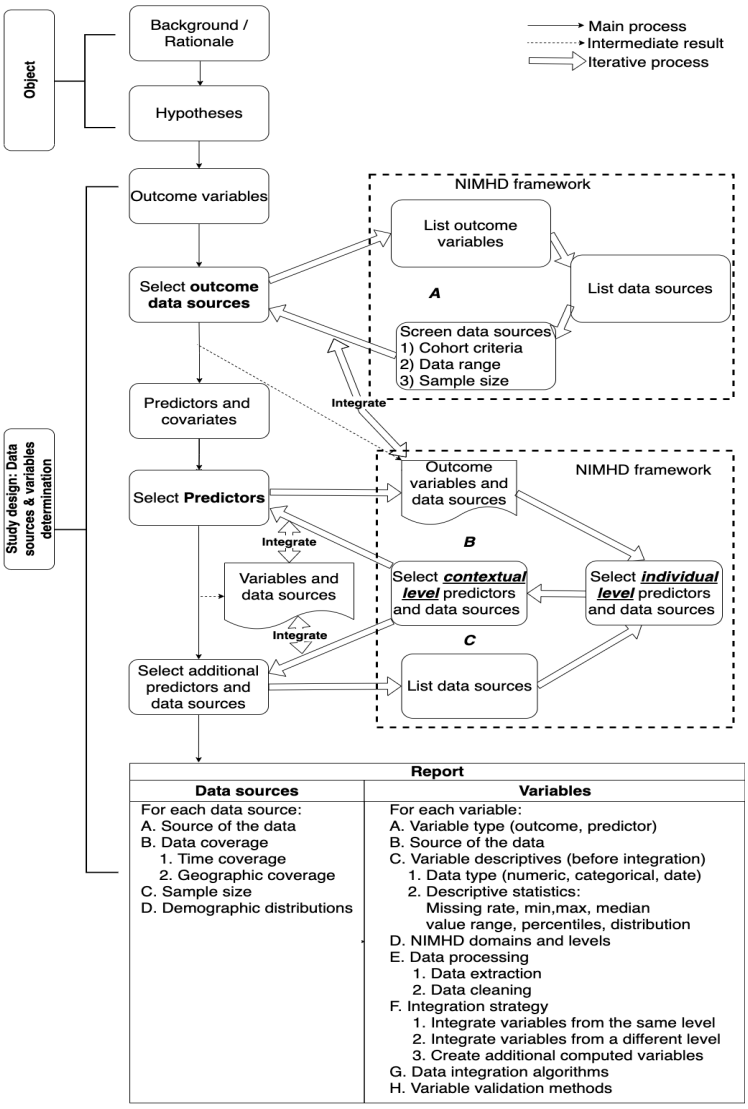

Figure 2. *An overview of the reporting guideline for RF variable and data source selection and data integration.*

391 ***Table 2.*** *ATTEST reporting guideline checklist*

	Item No	Recommendation	Page No
Objectives			
Background/rationale	1	Explain the scientific background and rationale for the study being reported in one or two sentences	
Prespecified hypotheses:	2	State prespecified hypotheses in on or two sentences	
Study design: data sources selection & variables selection & data integration			
Data sources	3a	Describe the time coverage	
	3b	Describe the geographic coverage	
	3c	Describe the sample size	
	3d	Describe the demographic distribution	
	3e	Describe the cohort criteria	
	3f	Describe the sources of biases (e.g., sample bias)	
	3g	Describe the data collection approach	
Dependent variables	4a	State the variable definition and variable type (e.g., primary outcome variable, secondary outcome variable)	
	4b	State the data source of dependent variable	
	4c	State the data type (e.g., numerical, categorical, date-time) of dependent variable	
	4d	State descriptive statistics (e.g., min, max. median, value range, percentile) of dependent variable	
	4e	State the NIMHD domains and levels of dependent variable	
Independent variables	5a	State the variable definition and variable type (e.g., primary predictor, secondary predictor)	
	5b	State the data source of dependent variable	
	5c	State the data type (e.g., numerical, categorical, date-time) of dependent variable	
	5b	State descriptive statistics (e.g., min, max. median, value range, percentile) of independent variable	
	5e	State the NIMHD domains and levels of independent variable	
Controlled variables	6a	State the variables type (e.g., numerical, categorical) of controlled variable	
	6b	State the data source of controlled variable	
	6c	State descriptive statistics (e.g., min, max. median, value range, percentile) of controlled variable	
	6d	State the NIMHD domains and levels of controlled variable	
Missing data	7a	For each data source, describe whether required or expected variable that is not present	
	7b	For each variable, describe method of how to handle missing data	
	7c	For each variable, describe the missing rate	
Data integration			
Data processing	8a	Data extraction: for each variable, describe how to process the raw data source to extract the variable	
	8b	Data cleaning: for each variable, describe the method used to detect and correct (or remove) the incorrect records, missing values or outliers	
Integration strategy	9	Describe the integration strategy for each variable:1) Integrate with variables from same level, 2) Integrate with variables from different levels, and 3) Creation of additional computed elements	

Integration algorithm	10	For each variable, describe the algorithm used to integrate it with variables from other data sources	
Variable validation	11	For each variable, describe data validation rule for the selected variable. Rule should identify both the variable and the validation algorithms	
Integrated variable	12	Describe the variable after integration and basic descriptive statistics (e.g., min, max. median, value range, percentile)	

*National Institute on Minority Health and Health Disparities (NIMHD)
Note: Please document the items for each data source and variable separately.

Development of the OD-ATTEST ontology

Based on the ATTEST reporting protocol above, we identified that 48 classes and 25 properties are needed in OD-ATTEST to represent the ATTEST reporting guideline. ***Figure 3*** shows the class hierarchy of OD-ATTEST. We reused classes from the following existing well-known ontologies: Ontology for Biomedical Investigations (OBI), Information Artifact Ontology (IAO), National Cancer Institute Thesaurus (NCIt), Statistics Ontology (STATO) and Semanticscience Integrated Ontology (SIO) as shown in ***Table 3***. Note that there are very few existing ontologies designed for the purpose of documenting the variable and data source selection and data integration process. The limited number of properties in these existing ontologies are not informative to represent the elements in the reporting guideline and their relationships, requiring us to create a large number of new properties in **OD-ATTEST.**

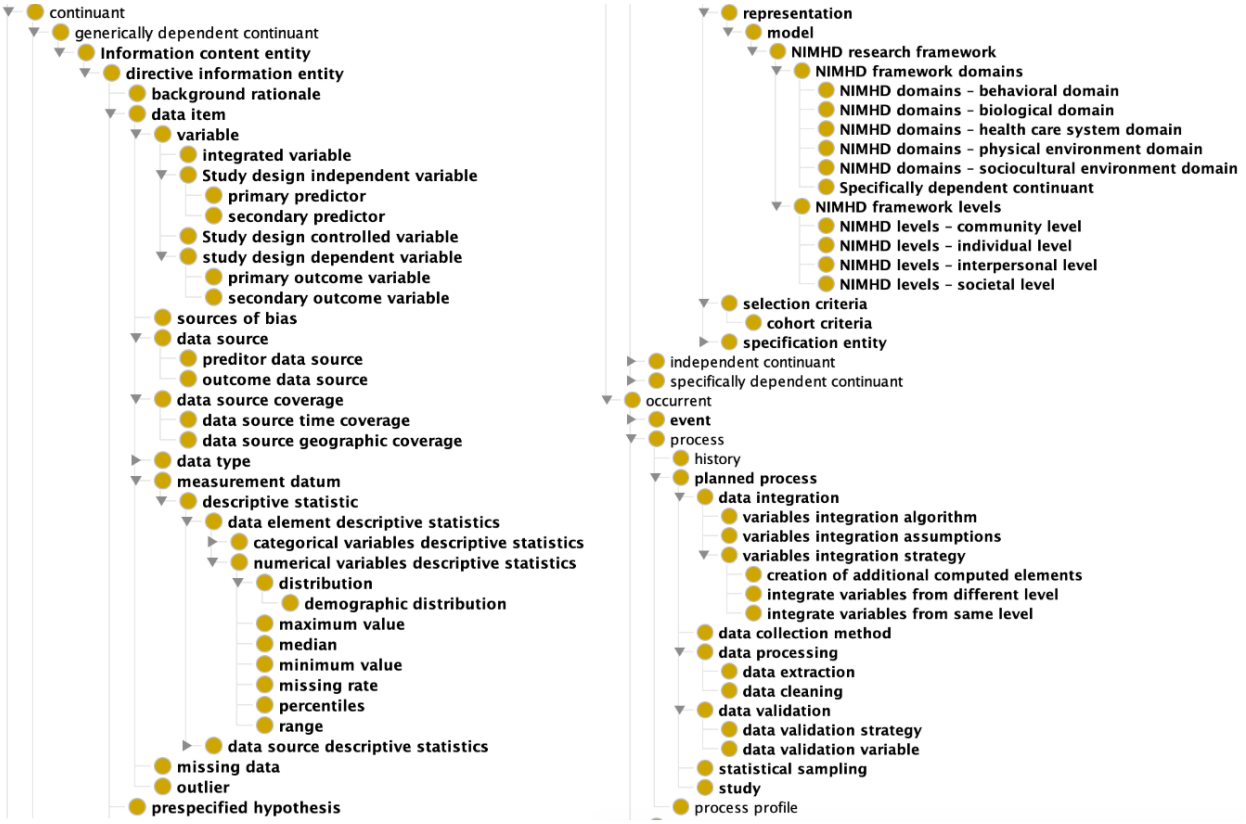

Figure 3. *The class hierarchy of OD-ATTEST.*

Table 3. The classes and properties reused or created for OD-ATTEST.

	Label	**Internationalized Resource Identifiers (IRIs)***	**Reference ontology**
Classes	objecitve	iao:0000005	IAO[1]
	data source	iao:0000100	
	measurement datum	iao:0000109	
	dependent variable	obi:0000751	OBI[2]
	independent variable	obi:0000750	
	controlled variable	obi:0000785	
	data processing	obi:0200000	
	study	ncit:C63536	NCIt[3]
	hypothesis	ncit:C28362	
	rationale	ncit:C80263	
	primary outcome	ncit:C142644	
	secondary outcome	ncit:C142680	
	sample size	ncit:C53190	
	missing data	ncit:C142610	
	data validation	ncit:C142500	
	data type	ncit:C42645	

	data collection method	ncit:C103159	
	data analysis	sio:001051	SIO[4]
	minimum value	stato:0000150	STATO[5]
	maximum value	stato:0000151	
	median	stato:0000574	
	mean	stato:0000573	
	value range	stato:0000035	
	percentile	stato:0000293	
	data distribution	stato:0000161	
	statistical sampling	stato:0000502	
	outlier	stato:0000036	
	primary predictor	od-attest:000015	OD-ATTEST[6]
	secondary predictor	od-attest:000016	
	demographic distribution	od-attest:000093	
	outcome variable data source	od-attest:000019	
	predictor data source	od-attest:000094	
	cohort criteria	od-attest:000008	
	descriptive statistic	od-attest:000012	
	missing rate	od-attest:000068	
	data source time coverage	od-attest:000023	
	data source geographic coverage	od-attest:000024	
	sources of bias	od-attest:000051	
	data integration	od-attest:000052	
	data extraction	od-attest:000054	
	data cleaning	od-attest:000055	
	integration strategy	od-attest:000056	
	integrate variables from same level	od-attest:000057	
	integrate variables from different levels	od-attest:000058	
	creation of additional elements	od-attest:000059	
	integration algorithm	od-attest:000060	
	validation strategy	od-attest:000068	
	integrated variable	od-attest:000096	
Properties	is determined by	od-attest:000097	OD-ATTEST
	has rationale	od-attest:000098	
	has objective	od-attest:000099	
	has data source	od-attest:000100	
	has cohort criteria	od-attest:000101	
	has demographic distribution	od-attest:000102	
	has sources of bias	od-attest:000103	
	has controlled variable	od-attest:000104	
	has independent variable	od-attest:000105	
	has dependent variable	od-attest:000106	
	has data type	od-attest:000107	
	has descriptive statistics	od-attest:000108	
	has NIMHD level	od-attest:000109	

	has NIMHD domain	od-attest:000110	
	has data collection approach	od-attest:000111	
	has sample size	od-attest:000112	
	has missing data	od-attest:000113	
	has data integration	od-attest:000114	
	has data processing	od-attest:000115	
	has data validation	od-attest:000116	
	has integration strategy	od-attest:000117	
	extracted from	od-attest:000118	
	has description	od-attest:000119	
	has time coverage	od-attest:000120	
	has geographic coverage	od-attest:000121	

*Prefix:
iao: <http://purl.obolibrary.org/obo/IAO_>
obi: <http://purl.obolibrary.org/obo/OBI_>
ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
sio: <http://purl.obolibrary.org/obo/SIO_>
stato: <http://purl.obolibrary.org/obo/STATO_>
od-attest: <http://purl.obolibrary.org/obo/OD-ATTEST/>
[1] Information Artifact Ontology
[2] Ontology for Biomedical Investigations
[3] National Cancer Institute Thesaurus
[4] Statistics Ontology
[5] Semanticscience Integrated Ontology
[6] Ontology for the Documentation of Variable and Data Source Selection and Integration Process

An OD-ATTEST-annotated report generated based on a mIDA case study following the reporting guideline

We annotated two of our previously published mIDA case studies: (1) one study that explored the impact of the relationships among socioeconomic status, individual smoking status, and community-level smoking rate on pharyngeal cancer survival [16], and (2) another study that created a semantic data integration framework to pool multi-level RFs from heterogenous data sources to support mIDA [25]. ***Table 4*** is the filled ATTEST checklist for the two studies. ***Figure 4*** shows a snippet of the ontology annotated variable and data source selection and integration process for the second study [25], while the corresponding semantic triples in RDF format using Turtle syntax is shown in ***Table 5.*** The items: RF variables, data sources, and data integration steps and their relationships are explicitly standardized and modeled using the classes and properties from OD-ATTEST.

420 ***Table 4***. An example of two previous mIDA case studies annotated using ATTEST checklist.

	Item No	Recommendation	Page No Study (1) [16]	Page No Study (2) [25]
Objectives				
Background/rationale	1	Explain the scientific background and rationale for the study being reported in one or two sentences	Page 1, section "*Abstract*", paragraph 1, line 1-7	Page 1, section "*Abstract*", paragraph 1, line 1-4
Prespecified hypotheses	2	State prespecified hypotheses in on or two sentences	Page 2, section "*Introduction*", paragraph 3, line 1-2	N/A
Study design: data sources selection & variables selection & data integration				
Data source	3a	Describe the time coverage	**FCDS**: Page 2, section "*Data source and case selection*", paragraph 1, line 2	**FCDS**: Page 4, section "Data sources", paragraph 1, line 11
			BRFSS: Page 2, section "*Data source and case selection*", paragraph 1, line 6	**BRFSS**: N/A
			2000 U.S. census data: Page 2, section *"Data source and case selection",* paragraph 1, line 7	**United States Census Bureau**: Page 4, section "Data sources", paragraph 1, line 23
				ATSDR: N/A
				County Health Ranking & Roadmaps: N/A
	3b	Describe the geographic coverage	**FCDS**: Page 2, section "*Data source and case selection*", paragraph 1, line 4-5"	**FCDS**: Page 4, section "Data sources", paragraph 1, line 12-14
			BRFSS: N/A	**BRFSS**: Page 10, section "Result", paragraph 2, line 7-8
			2000 U.S. census data: N/A	**United States Census Bureau**: N/A
				ATSDR: N/A
				County Health Ranking & Roadmaps: N/A
	3c	Describe the sample size	**FCDS**: Page 2, section "*Data source and case selection*", paragraph 2, line 7	**FCDS**: Page 4, section "Data sources", paragraph 2, line 6-7
			BRFSS: N/A	**BRFSS**: N/A
			2000 U.S. census data: N/A	**United States Census Bureau**: N/A
				ATSDR: N/A
				County Health Ranking & Roadmaps: N/A
	3d	Describe the demographic distribution	**FCDS**: Page 2, Table 1	N/A
			BRFSS: N/A	
			2000 U.S. census data: N/A	
	3e	Describe the Cohort criteria	**FCDS**: Page 2, section "*Data source and case selection*", paragraph 2, line 1-5	**FCDS**: Page 4, section "Data sources", paragraph 2, line 1-6
			BRFSS: N/A	**BRFSS**: N/A
			2000 U.S. census data: N/A	**United States Census Bureau**: N/A
				ATSDR: N/A
				County Health Ranking & Roadmaps: N/A
	3f	Describe the sources of bias	N/A	N/A
	3g	Describe the data collection approach	N/A	**FCDS**: N/A
				BRFSS: Page 4, section "Data sources", paragraph 2, line 6-7
				United States Census Bureau: N/A
				ATSDR: N/A
				County Health Ranking & Roadmaps: N/A
Dependent variable	4a	State the variable definition and variable type (e.g., primary outcome	**Survival time:** Page 2, section "*Variable definitions*", line 1-3	**Cancer survival**: Page 4, section "Data integration use case: The multi-level integrative data analysis of Cancer survival", paragraph 1, line 1-2

		variable, secondary outcome variable)		
	4b	State the data source of dependent variable	**Survival time:** Page 2, section "*Data source and case selection*", paragraph 1, line 2	**Cancer survival**: Page 4, section "Data sources", paragraph 1, line 9-14
	4c	State the data type (e.g., numerical, categorical, date-time) of dependent variable	**Survival time:** Page 2, section "*Variable definitions*", paragraph 1, line 1	**Cancer survival**: N/A
	4d	State descriptive statistics (e.g., min, max. median, value range, percentile) of dependent variable	**Survival time:** Page 4, Table 1	Cancer survival: N/A
	4e	State the NIMHD domain and levels of dependent variable	**Survival time:** Page 2, section "*Data source and case selection*", paragraph 1, line 1-2	**Cancer survival**: Page 4, section "Data sources", paragraph 2, line 15
Independent variable	5a	State the variable definition and variable type (e.g., primary predictor, secondary predictor)	**Socioeconomic status:** Page 2, section "*Variable definitions*", paragraph 3, line 1-2	**Demographic variables**: Page 5, table 1
			Individual smoking: Page 2, section "*Data source and case selection*", paragraph 2, line 1-2	**Smoking status:** Page 10, section "The ontology for Cancer research variables (OCRV)", paragraph 2, line 13- 27
			Regional smoking: Page 3, section "*Data source and case selection*", paragraph 2, line 4-6	**Marital status:** Page 14, section "Type 4: Queries that generate results based on the knowledge encoded in ontology", paragraph 2, line 7- 10
				Insurance payer: Page 5, table 1
				Residency: Page 5, table 1
				Age at diagnosis: Page 5, table 1
				Year of diagnosis: Page 5, table 1
				Tumor stage: Page 5, table 1
				Tumor type: Page 5, table 1
				Treatment procedure: Page 5, table 1
				Census Tract SVI: Page 14, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 5-16
				Census tract high school completion rates: Page 5, table 1
				Census tract family poverty rates: Page 5, table 1
				Census tract rurality status: Page 4, section "Data integration use case: The multi-level integrative data analysis of Cancer survival", paragraph 1, line 8-11
				County adult mental and physical health status: Page 5, table 1
				County density of primary care physicians: Page 5, table 1
				County smoking rate: Page 10, section "The ontology for Cancer research variables (OCRV)", paragraph 2
				County alcohol consumption rate: Page 5, table 1
	5b	State the data type (e.g., numerical, categorical) of independent variable	**Socioeconomic status:** Page 2, section "*Variable definitions*", paragraph 3, line 9-10	**Demographic variables**: N/A
			Individual smoking: Page 2, section "*Data source and case selection*", paragraph 2, line 2-3	**Smoking status:** Page 13, table 3
			Regional smoking: Page 3, section "*Data source and case selection*", paragraph 2, line 4-6	**Marital status:** Page 14, section "Type 4: Queries that generate results based on the knowledge encoded in ontology", paragraph 2, line 7- 10

				Insurance payer: N/A
				Residency: N/A
				Age at diagnosis: Page 16, figure 6
				Year of diagnosis: Page 16, figure 6
				Tumor stage: N/A
				Tumor type: Page 4, section "Data sources", paragraph 2, line 1-6
				Treatment procedure: Page 5, table 1
				Census Tract SVI: Page 14, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 5-16
				Census tract high school completion rates: N/A
				Census tract family poverty rates: N/A
				Census tract rurality status: N/A
				County adult mental and physical health status: N/A
				County density of primary care physicians: N/A
				County smoking rate: Page 10, section "The ontology for Cancer research variables (OCRV)", paragraph 2
				County alcohol consumption rate: N/A
	5c	State the data source of independent variable	**Socioeconomic status:** Page 2, section "*Data source and case selection*", paragraph 1, line 6-7	Page 5, table 1
			Individual smoking: Page 2, section "*Data source and case selection*", paragraph 1, line 1-2	
			Regional smoking: Page 2, section *"Data source and case selection"*, paragraph 1, line 7-10	
	5d	State descriptive statistics (e.g., min, max. median, value range, percentile) of independent variable	Page 4, table 1	N/A
	5e	State the NIMHD domain and levels of independent variable	**Socioeconomic status:** Page 2, section "*Data source and case selection*", paragraph 1, line 6	Page 5, table 1
			Individual smoking: Page 2, section "*Data source and case selection*", paragraph 2, line 1	
			Regional smoking: Page 3, section "*Data source and case selection*", paragraph 2, line 4-6	
Controlled variable	6a	State the controlled variable and variable type (e.g., numerical, categorical) of controlled variable	**Age of diagnosis:** Page 2, section "*Variable definitions*", paragraph 1, line 10-13	N/A
			Anatomic site: Page 2, section "*Variable definitions*", paragraph 1, line 2-9	
			Race-ethnicity: Page 4, table 1	
			Marital status: Page 4, table 1	
			Insurance: Page 4, table 1	
			Year of diagnosis: Page 4, table 1	
			Gender: Page 4, table 1	
			Stage of diagnosis: Page 4, table 1	
			Treatment: Page 4, table 1	
	6b	State the data source of controlled variable	Page 2, section "*Data source and case selection*", paragraph 1, line 2*	N/A

	6c	State descriptive statistics (e.g., min, max. median, value range, percentile) of controlled variable	Page 2, section "*Data source and case selection*", paragraph 1, line 2*	N/A
	6d	State the NIMHD domain and levels of controlled variable	Page 2, section "*Data source and case selection*", paragraph 1, line 1-5*	N/A
Missing data	7a	For each data source, describe whether required or expected variable that is not present	N/A	N/A
	7b	For each variable, describe method of how to handle missing data	N/A	N/A
	7c	For each variable, describe the missing rate	N/A	N/A
Data processing	9a	Data extraction: for each variable, describe how to process the raw data source to extract the variable	N/A	**Demographic variables**: Page 15, figure 5
				Age at diagnosis: Page 16, figure 6
				Census Tract SVI: Page 16, figure 7
				County smoking rate: Page 17, figure 8
				Marital status: Page 18, figure 9
	9b	Data cleaning: for each variable, describe the method used to detect and correct (or remove) the incorrect records, missing values or outliers	N/A	N/A
Integration strategy	10	Describe the integration strategy for each variable:1) Integrate with variables from same level, 2) Integrate with variables from different levels, and 3) Creation of additional computed elements	**Socioeconomic status:** Page 2, section "*Variable definitions*", paragraph 3, line 6-7.	**Demographic variables**: Page 15, figure 5
			Regional smoking: Page 2, section "*Variable definitions*", paragraph 2, line 4-5.	**Age at diagnosis:** Page 16, figure 6
				Census Tract SVI: Page 16, figure 7
				County smoking rate: Page 17, figure 8
				Marital status: Page 18, figure 9
				Census tract high school completion rates: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				Census tract family poverty rates: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				Census tract rurality status: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				County adult mental and physical health status: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3

				County density of primary care physicians: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				County alcohol consumption rate: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
Integration algorithms	11	For each variable, describe the algorithm used to integrate it with variables from other data sources	N/A	**Demographic variables**: Page 15, figure 5
				Age at diagnosis: Page 16, figure 6
				Census Tract SVI: Page 16, figure 7
				County smoking rate: Page 17, figure 8
				Marital status: Page 18, figure 9
				Census tract high school completion rates: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				Census tract family poverty rates: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				Census tract rurality status: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				County adult mental and physical health status: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				County density of primary care physicians: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
				County alcohol consumption rate: Page 15, section "Type 3: Queries that are used to link a patient to contextual factors through geographic variables", paragraph 1, line 1- 3
Variable validation	12	For each variable, describe data validation rule for the selected variable. Rule should identify both the variable and the validation algorithms	N/A	**Demographic variables**: Page 19, section "Data quality and consistency checks of the source data using the ontology"
Integrated variable	13	Describe the variable after integration and basic descriptive statistics (e.g., min, max. median, value range, percentile)	N/A	Page 18, Table 4

*Note: if the reported items for all variables or data sources are described at the same place, you can list the page/section/table information at once. For the integration related items, we only presented variables that have the information (N/A will not be showed in the table).
FCDS: Florida Cancer Data System
ATSDR: Agency for Toxic Substances& Disease Registry
BRFSS: Behavioral Risk Factor Surveillance System

Figure 4. *An OD-ATTEST-annotated report generated based on a mIDA case study.*

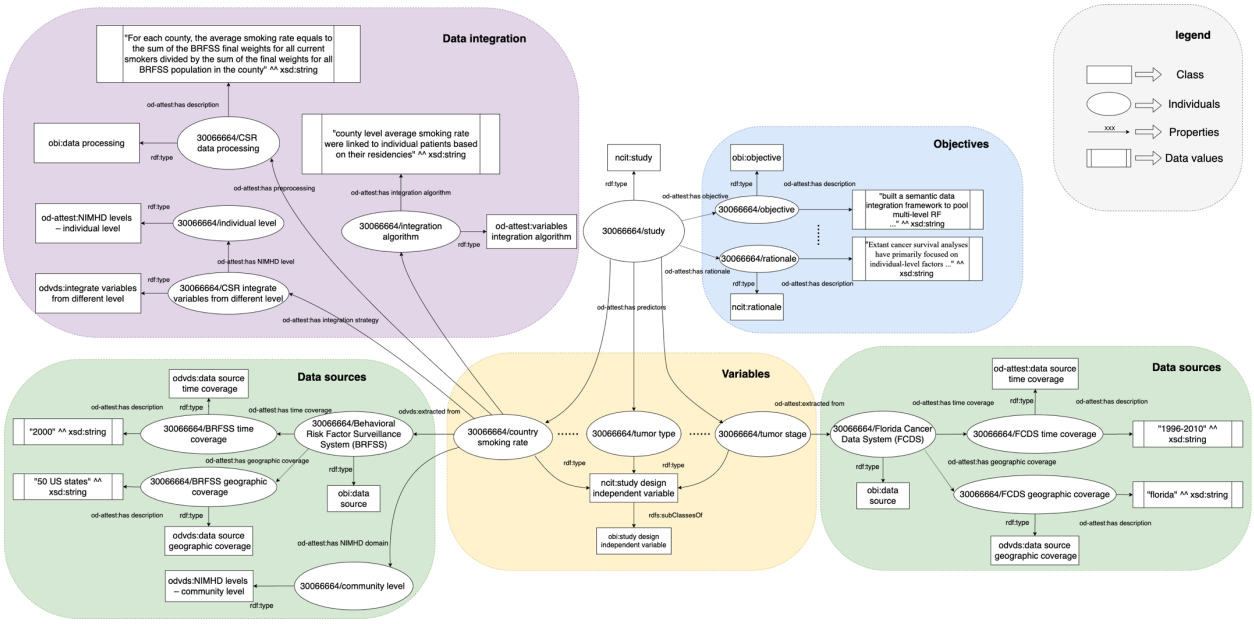

Table 5. *An example of annotated semantic triples represented in RDF format using Turtle syntax.*

Prefix	@prefix od-attest: <http://www.semanticweb.org/od-attest#>. @prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>. @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>. @prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
RDF[1] triples	od-attest:30066664 rdf:type ncit:study; od-attest:has rationale od-attest:30066664/rationale; od-attest:has objective od-attest:30066664/objective. od-attest:30066664/rationale rdf:type ncit:rationale; od-attest:has description "Extant cancer survival analyses have..." ^^ xsd:string. od-attest:30066664/objective rdf:type ncit:objective; od-attest:has description "built a semantic data integration …" ^^ xsd:string.
[1] Resource Description Framework	

DISCUSSION

In this study, we first developed a reporting guideline, ATTEST, to provide a theory-driven approach to guide the RF variable and data source selection and integration process in cancer outcomes research. We then proposed an ontology-based approach to annotate the items in our reporting guideline so that information relevant to variables, data sources and data integration in mIDA studies can be explicitly documented. To develop the reporting guideline, we conducted a systematic search to identify useful reporting items to improve our selection and data integration process. We categorized these reporting guidelines based on their reported data source domains and levels according to NIMHD framework, so that we can identify items need to be reported when selecting variables or data sources from different domains and levels. For example, when report population-level estimates (variables) [42], the information regarding the sources of bias (e.g., selection bias) need to be documented. Therefore, we updated our previous reporting guideline and added "*sources of bias*" as a reporting item when documenting data sources. This is important, because subsequent data processing steps might be needed to correct the bias. Further,

The use of NIMHD framework can also help researchers to systematically think and structure the variable and data source selection process when considering multi-level RF variables from heterogenous data sources. For example, if an investigator is considering smoking related risk factors in cancer outcomes research, following the NIMHD framework, one can start with variables in the behavioral domain and then list potential smoking related variables for each level of influences step by step, such as individual smoking status at the individual level, second hand smoke exposure at the interpersonal level, county level smoking rate at the community level, and

451 smoking policies or laws (e.g., federal minimum age to purchase tobacco products) at the
452 societal level. The same process can be applied to select other smoking related variables from
453 other domains of influences. In this way, investigators can systematically think and evaluate the
454 confounding effects and cross-level interactions among those selected variables which are
455 usually ignored in previous cancer outcome studies using a single data source.
456
457 We provided a ATTEST checklist (1) to help researchers clearly document each step of their RF
458 and data source selection and integration process, and (2) to improve the completeness and
459 transparency of their mIDA studies. As shown in ***Table 4,*** we used the ATTEST checklist to
460 report two previous mIDA studies. Based on the checklist, we can easily (1) check whether
461 these mIDA studies document required items that can help other researchers replicate their
462 studies, and (2) compare their variables, data sources and data integration processes. As shown
463 in ***Table 4,***, we found that there are 3 items never discussed in either of the two studies including
464 "*sources of bias"*, "*missing data"* for selected variables, and "*data cleaning*" (i.e., method used
465 to detect and correct or remove the incorrect records, missing values or outliers). All three items
466 are relevant to data quality issues, where rarely being discussed or documented in these mIDA
467 studies or even more broadly in cancer outcomes research. Nevertheless, data quality issues such
468 as missing data can dramatically affect the results of the cancer outcomes research (e.g., in
469 cancer survival prediction) [60]. Comparing the two case mIDA studies, the data integration
470 process was not well-documented in the first study [16], where most of the items relevant to data
471 integration are blank; while, in the other study [25], the processes about data processing, data
472 integration, and data validation were all clearly documented according to the ATTEST checklist.

Therefore, using this checklist, one can improve the completeness of their documentation on the selection and integration process as shown in ***Table 4***.

The OD-ATTEST ontology provides a way to standardize the documentation of the mIDA study process from variable and data source selection to data integration. Also, the ontology-based annotations of the report is beneficial because it provides an initial step towards a report that is not only readable and understandable by human but also potentially executable by machines. After transforming these annotations into semantic triples, the report can be stored into a knowledge base and represented as knowledge graphs (***Figure 4***) to facilitate examination and analysis of these mIDA reports, enabling robust sharing and comparison of different mIDA studies.

Limitations and future work

Most of the reporting guidelines we reviewed from the EQUATOR network have limited information on how to document the data integration process, indicating a significant gap in existing practice. Nevertheless, we were able to summarize the key elements need to be reported for the integration process based on 3 existing guidelines and our own previous experience on semantic data integration case studies. As a future study, one shall conduct a systematic review on data integration literatures to summarize relevant reporting items to improve the reporting guideline. Meanwhile, we will conduct a yearly review of existing reporting guidelines following the reviewing process discussed in ***Figure 1*** to identify new reporting items of interest and keep our framework up to date. Further, beyond standardized reporting, our ultimate goal is to let computers understand the ontology-annotated report (in RDF triples) regarding (1) how

different variables are defined and represented and (2) how different variables are selected and integrated, so that machines can automatically repeat these processes and generate integrated dataset based on an executable ontology-annotated report. For variable definition and representation, it is important to recognize and being interoperable with existing data standards and common data models (CDM) such as those that standardized exchanging of EHRs data including the national Patient-Centered Clinical Research Network (PCORnet) CDM, the Observational Medical Outcomes Partnership (OMOP) from the Observational Health Data Sciences and Informatics (OHDSI) network, and the uprising Fast Healthcare Interoperability Resources (FHIR) protocol adopted by major EHR system vendors. Developing the ontology against these CDMs that have already standardized existing data resources would be critical to assure the generalizability of our framework. Nevertheless, for modeling the variable selection and integration processes as shown in ***Figure 4***, more fine-grained information regarding the variables, data sources and the integration process are currently documented as free-text descriptions. We face challenges in transforming these "free-text" information into executable algorithms (e.g., a data processing step that calculates BMI using weight and height). Such information is related to the concept of data provenance—"*a type of metadata, concerned with the history of data, its origin and changes made to it*" [61]. The importance of data provenance is widely recognized, especially for study reproducibility and replicability. More than one-half of the systematic efforts to reproduce computational results across different fields have failed, mainly due to insufficient detail on digital artifacts, such as data, code, and computational workflow [62]. However, descriptions of data provenance are often neglected or inadequate in scientific literature due to the lack of a tractable, easily operated approach with supporting tools. Future studies that focus on the development of easy-to-use tools with a standardized framework

519 to persist end-to-end data provenance with high integrity including intermediate processes and
520 data products are urgently needed. Further, future developments of tools and platforms to
521 automate the documentation process, where the data elements and associated information (e.g.,
522 levels and domains) are also automatically annotated with the standardized ontology are
523 warranted.
524

525 ## CONCLUSIONS

526 In this paper, we have proposed and developed an ontology-based reporting guideline solving
527 some key challenges in current mIDA studies for cancer outcomes research, through providing
528 (1) a theory-driven guidance for multi-level and multi-domain RF variable and data source
529 selection; and (2) a standardized documentation of the data selection and integration processes
530 powered by an ontology, thus a way to enable sharing of mIDA study reports among researchers.
531

532 **List of abbreviations**

533	ACS	American Cancer Society
534	BRFSS	Behavioral Risk Factor Surveillance System
535	EQUATOR	Enhancing the QUAlity and Transparency Of health Research
536	FCDS	Florida Cancer Data System
537	mIDA	Multi-level Integrative Data Analysis
538	NIH	National Institute of Health
539	NIMHD	Minority Health and Health Disparities
540	OD-ATTEST	Ontology for the Documentation of Variable and Data Source Selection and
541	Integration Process	

RF	Risk Factor
US	United States
RUCA	Rural-Urban Commuting Area
NCHS	National Center for Health Statistics
BFO	Basic Formal Ontology
NCBO	National Center for Biomedical Ontology
RDF	Resource Description Framework
GRIPS	Genetic RIsk Prediction Studies
COHERE	Checklist for One Health Epidemiological Reporting of Evidence
EHR	Electronic Health Records
OBI	Ontology for Biomedical Investigations
IAO	Information Artifact Ontology
NCIt	National Cancer Institute Thesaurus
STATO	Statistics Ontology
SIO	Semanticscience Integrated Ontology
CDM	Common Data Model
PCORnet	The national Patient-Centered Clinical Research Network

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publish

Not applicable

Availability of data and materials

The reviewed reporting guidelines are available in the public: Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network (https://www.equator-network.org/reporting-guidelines/).

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported in part by the National Institute of Health (NIH) awards UL1TR001427 and R01CA246418 and Patient-Centered Outcomes Research Institute (PCORI) award ME-2018C3-14754. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.

Authors' contributions

The work presented here was carried out in collaboration among all authors. YG and JB designed the study. YG, QL and HZ were involved in acquisition of the data and review of existing reporting guidelines. HZ wrote the initial draft of the manuscript with substantial support from YG and JB. YG and JB provided expert opinion during the ontology curation process and guided the design of the ontology. All authors provided critical feedback on the

study design, reviewed and edited the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

None

References

1. World Health Organization. Cancer - key facts. 2018. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed 2 Jan 2020.

2. Atlanta: American Cancer Society. Cancer Facts & Figures 2019. 2019. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2019.html. Accessed 2 Jan 2020.

3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69:7–34. doi:10.3322/caac.21551.

4. National Cancer Institute. Cancer Risk Factors. https://training.seer.cancer.gov/disease/cancer/risk.html. Accessed 2 Jan 2020.

5. Andrew AS, Parker S, Anderson JC, Rees JR, Robinson C, Riddle B, et al. Risk Factors for Diagnosis of Colorectal Cancer at a Late Stage: a Population-Based Study. J Gen Intern Med. 2018;33:2100–5. doi:10.1007/s11606-018-4648-7.

6. Mobley LR, Kuo T-M. Demographic Disparities in Late-Stage Diagnosis of Breast and Colorectal Cancers Across the USA. J Racial Ethn Health Disparities. 2017;4:201–12. doi:10.1007/s40615-016-0219-y.

7. Markossian TW, Hines RB. Disparities in late stage diagnosis, treatment, and breast cancer-related death by race, age, and rural residence among women in Georgia. Women Health. 2012;52:317–35.

8. Chatterjee NA, He Y, Keating NL. Racial differences in breast cancer stage at diagnosis in the mammography era. Am J Public Health. 2013;103:170–6.

9. Montealegre JR, Zhou R, Amirian ES, Follen M, Scheurer ME. Nativity disparities in late-stage diagnosis and cause-specific survival among Hispanic women with invasive cervical cancer: an analysis of Surveillance, Epidemiology, and End Results data. Cancer Causes Control. 2013;24:1985–94. doi:10.1007/s10552-013-0274-1.

618 10. Baquet CR, Mishra SI, Commiskey P, Ellison GL, DeShields M. Breast cancer epidemiology in
619 blacks and whites: disparities in incidence, mortality, survival rates and histology. J Natl Med
620 Assoc. 2008;100:480–8.

621 11. Yasmeen S, Xing G, Morris C, Chlebowski RT, Romano PS. Comorbidities and mammography
622 use interact to explain racial/ethnic disparities in breast cancer stage at diagnosis. Cancer.
623 2011;117:3252–61.

624 12. Echeverría SE, Borrell LN, Brown D, Rhoads G. A local area analysis of racial, ethnic, and
625 neighborhood disparities in breast cancer staging. Cancer Epidemiol Biomark Prev Publ Am
626 Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2009;18:3024–9.

627 13. NIMHD. NIMHD Research Framework. https://www.nimhd.nih.gov/about/
628 overview/research-framework.html. Accessed 28 Jun 2019.

629 14. Dahlberg LL, Krug EG. Violence a global public health problem. Ciênc Saúde Coletiva.
630 2006;11:277–92. doi:10.1590/S1413-81232006000200007.

631 15. Keegan TH, Quach T, Shema S, Glaser SL, Gomez SL. The influence of nativity and
632 neighborhoods on breast cancer stage at diagnosis and survival among California Hispanic
633 women. BMC Cancer. 2010;10:603. doi:10.1186/1471-2407-10-603.

634 16. Guo Y, Logan HL, Marks JG, Shenkman EA. The relationships among individual and regional
635 smoking, socioeconomic status, and oral and pharyngeal cancer survival: a mediation analysis.
636 Cancer Med. 2015;4:1612–9.

637 17. Giordano A. Data integration blueprint and modeling: techniques for a scalable and
638 sustainable architecture. Upper Saddle River, NJ: IBM Press Pearson; 2011.

639 18. Schloss PD. Identifying and Overcoming Threats to Reproducibility, Replicability,
640 Robustness, and Generalizability in Microbiome Research. mBio. 2018;9:e00525-18,
641 /mbio/9/3/mBio.00525-18.atom. doi:10.1128/mBio.00525-18.

642 19. Alonso-Calvo R, Paraiso-Medina S, Perez-Rey D, Alonso-Oset E, van Stiphout R, Yu S, et al. A
643 semantic interoperability approach to support integration of gene expression and clinical data
644 in breast cancer. Comput Biol Med. 2017;87:179–86. doi:10.1016/j.compbiomed.2017.06.005.

645 20. Kondylakis H, Claerhout B, Keyur M, Koumakis L, van Leeuwen J, Marias K, et al. The
646 INTEGRATE project: Delivering solutions for efficient multi-centric clinical research and trials. J
647 Biomed Inform. 2016;62:32–47. doi:10.1016/j.jbi.2016.05.006.

648 21. METABRIC Group, Papatheodorou I, Crichton C, Morris L, Maccallum P, Davies J, et al. A
649 metadata approach for clinical data management in translational genomics studies in breast
650 cancer. BMC Med Genomics. 2009;2. doi:10.1186/1755-8794-2-66.

651 22. Centre for Statistics in Medicine, NDORMS, University of Oxford. Enhancing the QUAlity and
652 Transparency Of health Research. https://www.equator-network.org/reporting-guidelines/.
653 Accessed 28 Jan 2020.

654 23. Zhang H, Guo Y, Bian J. Ontology for Documentation of Variable and Data Source Selection
655 Process to Support Integrative Data Analysis in Cancer Outcomes Research. In: SEPDA@ISWC.
656 2019.

657 24. Guo Y, Bian J, Modave F, Li Q, George TJ, Prosperi M, et al. Assessing the effect of data
658 integration on predictive ability of cancer survival models. Health Informatics J.
659 2019;:1460458218824692.

660 25. Zhang H, Guo Y, Li Q, George TJ, Shenkman E, Modave F, et al. An ontology-guided semantic
661 data integration framework to support integrative data analysis of cancer survival. BMC Med
662 Inform Decis Mak. 2018;18. doi:10.1186/s12911-018-0636-4.

663 26. Rural-Urban Commuting Area Codes. 2019. https://www.ers.usda.gov/data-products/rural-
664 urban-commuting-area-codes.aspx. Accessed 28 Jan 2020.

665 27. National Center for Health Statistics, Office of Analysis and Epidemiology. NCHS Urban-Rural
666 Classification Scheme for Counties. 2017.
667 https://www.cdc.gov/nchs/data_access/urban_rural.htm#2013_Urban-
668 Rural_Classification_Scheme_for_Counties. Accessed 28 Jan 2017.

669 28. Arp R, Smith B, Spear AD. Building Ontologies With Basic Formal Ontology. The MIT Press;
670 2015. doi:10.7551/mitpress/9780262527811.001.0001.

671 29. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal:
672 enhanced functionality via new Web services from the National Center for Biomedical Ontology
673 to access and use ontologies in software applications. Nucleic Acids Res. 2011;39 Web Server
674 issue:W541-545.

675 30. David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, Gavin Carothers, Lex Machina. RDF
676 1.1 Turtle. 2014. https://www.w3.org/TR/2014/REC-turtle-20140225/Overview.html. Accessed
677 31 Jan 2020.

678 31. Janssens ACJW, Ioannidis JPA, van Duijn CM, Little J, Khoury MJ, GRIPS Group. Strengthening
679 the reporting of Genetic RIsk Prediction Studies: the GRIPS Statement. PLoS Med.
680 2011;8:e1000420.

681 32. Davis MF, Rankin SC, Schurer JM, Cole S, Conti L, Rabinowitz P, et al. Checklist for One
682 Health Epidemiological Reporting of Evidence (COHERE). One Health. 2017;4:14–21.
683 doi:10.1016/j.onehlt.2017.07.001.

684 33. Leech NL, Onwuegbuzie AJ. Guidelines for Conducting and Reporting Mixed Research in the
685 Field of Counseling and Beyond. J Couns Dev. 2010;88:61–9. doi:10.1002/j.1556-
686 6678.2010.tb00151.x.

687 34. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable
688 prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann
689 Intern Med. 2015;162:55. doi:10.7326/M14-0697.

690 35. Kerr KF, Meisner A, Thiessen-Philbrook H, Coca SG, Parikh CR. RiGoR: reporting guidelines to
691 address common sources of bias in risk model development. Biomark Res. 2015;3:2.

692 36. Jason LA, Unger ER, Dimitrakoff JD, Fagin AP, Houghton M, Cook DB, et al. Minimum data
693 elements for research reports on CFS. Brain Behav Immun. 2012;26:401–6.

694 37. Fitchett EJA, Seale AC, Vergnano S, Sharland M, Heath PT, Saha SK, et al. Strengthening the
695 Reporting of Observational Studies in Epidemiology for Newborn Infection (STROBE-NI): an
696 extension of the STROBE statement for neonatal infection research. Lancet Infect Dis.
697 2016;16:e202–13.

698 38. White RG, Hakim AJ, Salganik MJ, Spiller MW, Johnston LG, Kerr L, et al. Strengthening the
699 Reporting of Observational Studies in Epidemiology for respondent-driven sampling studies:
700 "STROBE-RDS" statement. J Clin Epidemiol. 2015;68:1463–71.

701 39. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The
702 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement:
703 guidelines for reporting observational studies. Ann Intern Med. 2007;147:573–7.

704 40. Jackson DL. Reporting results of latent growth modeling and multilevel modeling analyses:
705 some recommendations for rehabilitation psychology. Rehabil Psychol. 2010;55:272–85.

706 41. Wolfe F, Lassere M, van der Heijde D, Stucki G, Suarez-Almazor M, Pincus T, et al.
707 Preliminary core set of domains and reporting requirements for longitudinal observational
708 studies in rheumatology. J Rheumatol. 1999;26:484–9.

709 42. Stevens GA, Alkema L, Black RE, Boerma JT, Collins GS, Ezzati M, et al. Guidelines for
710 Accurate and Transparent Health Estimates Reporting: the GATHER statement. The Lancet.
711 2016;388:e19–23. doi:10.1016/S0140-6736(16)30388-9.

712 43. Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, von Elm E, et al. STrengthening the
713 REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement.
714 PLoS Med. 2009;6:e22.

715 44. Hollenbach JA, Mack SJ, Gourraud P-A, Single RM, Maiers M, Middleton D, et al. A
716 community standard for immunogenomic data reporting and analysis: proposal for a
717 STrengthening the REporting of Immunogenomic Studies statement. Tissue Antigens.
718 2011;78:333–44.

719 45. Field N, Cohen T, Struelens MJ, Palm D, Cookson B, Glynn JR, et al. Strengthening the
720 Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the
721 STROBE statement. Lancet Infect Dis. 2014;14:341–52.

722 46. Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JPA, Kirsch-Volders M, et al.
723 STrengthening the Reporting of OBservational studies in Epidemiology - Molecular
724 Epidemiology (STROBE-ME): an extension of the STROBE statement. Eur J Clin Invest.
725 2012;42:1–16.

726 47. Dixon WG, Carmona L, Finckh A, Hetland ML, Kvien TK, Landewe R, et al. EULAR points to
727 consider when establishing, analysing and reporting safety data of biologics registers in
728 rheumatology. Ann Rheum Dis. 2010;69:1596–602.

729 48. Zavada J, Dixon WG, Askling J, EULAR Study group on Longitudinal Observational Registers
730 and Drug Studies. Launch of a checklist for reporting longitudinal observational drug studies in
731 rheumatology: a EULAR extension of STROBE guidelines based on experience from biologics
732 registries. Ann Rheum Dis. 2014;73:628.

733 49. Singh JP, Yang S, Mulvey EP, RAGEE Group. Reporting guidance for violence risk assessment
734 predictive validity studies: the RAGEE Statement. Law Hum Behav. 2015;39:15–22.

735 50. Lachat C, Hawwash D, Ocké MC, Berg C, Forsum E, Hörnell A, et al. Strengthening the
736 Reporting of Observational Studies in Epidemiology—Nutritional Epidemiology (STROBE-nut):
737 An Extension of the STROBE Statement. PLOS Med. 2016;13:e1002036.
738 doi:10.1371/journal.pmed.1002036.

739 51. De Geest S, Zullig LL, Dunbar-Jacob J, Helmy R, Hughes DA, Wilson IB, et al. ESPACOMP
740 Medication Adherence Reporting Guideline (EMERGE). Ann Intern Med. 2018;169:30–5.

741 52. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, et al. Reporting to
742 Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies
743 V1.0. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2017;20:1009–22.

744 53. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting
745 of data quality in distributed data networks. EGEMS Wash DC. 2015;3:1052.

746 54. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of
747 studies conducted using observational routinely collected health data statement for
748 pharmacoepidemiology (RECORD-PE). BMJ. 2018;:k3532. doi:10.1136/bmj.k3532.

749 55. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of
750 studies Conducted using Observational Routinely-collected health Data (RECORD) Statement.
751 PLOS Med. 2015;12:e1001885. doi:10.1371/journal.pmed.1001885.

752 56. Bennett DA, Brayne C, Feigin VL, Barker-Collo S, Brainin M, Davis D, et al. Development of
753 the Standards of Reporting of Neurological Disorders (STROND) checklist: A guideline for the

754 reporting of incidence and prevalence studies in neuroepidemiology. Neurology. 2015;85:821–
755 8.

756 57. Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative
757 effectiveness research: defining, reporting and interpreting nonrandomized studies of
758 treatment effects using secondary data sources: the ISPOR Good Research Practices for
759 Retrospective Database Analysis Task Force Report--Part I. Value Health J Int Soc
760 Pharmacoeconomics Outcomes Res. 2009;12:1044–52.

761 58. Holtfreter B, Albandar JM, Dietrich T, Dye BA, Eaton KA, Eke PI, et al. Standards for reporting
762 chronic periodontitis prevalence and severity in epidemiologic studies: Proposed standards
763 from the Joint EU/USA Periodontal Epidemiology Working Group. J Clin Periodontol.
764 2015;42:407–12.

765 59. Tacconelli E, Cataldo MA, Paul M, Leibovici L, Kluytmans J, Schröder W, et al. STROBE-AMS:
766 recommendations to optimise reporting of epidemiological studies on antimicrobial resistance
767 and informing improvement in antimicrobial stewardship. BMJ Open. 2016;6:e010134.
768 doi:10.1136/bmjopen-2015-010134.

769 60. Barakat MS, Field M, Ghose A, Stirling D, Holloway L, Vinod S, et al. The effect of imputing
770 missing clinical attribute values on training lung cancer survival prediction model performance.
771 Health Inf Sci Syst. 2017;5:16.

772 61. Glavic B, Dittrich KR. Data Provenance: A Categorization of Existing Approaches. In:
773 Datenbanksysteme in Business, Technologie und Web (BTW). Aachen: Ges. für Informatik;
774 2007. p. 227–41.

775 62. Committee on Reproducibility and Replicability in Science, Board on Behavioral, Cognitive,
776 and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social
777 Sciences and Education, Nuclear and Radiation Studies Board, Division on Earth and Life
778 Studies, et al. Reproducibility and Replicability in Science. Washington, D.C.: National
779 Academies Press; 2019. doi:10.17226/25303.

780

781