NEST
COORDINATING CENTER
An Initiative of MDIC

1655 Fort Myer Drive, Suite 1200
Arlington, VA 22209

202.828.1600
NESTcc.org

THE NATIONAL EVALUATION CENTER FOR HEALTH TECHNOLOGY

Appendix. Data Quality Reporting Checklist

#	Component/Question	Response/Details
Section 1: Dataset Overview		
1	Data owner and contact information	
3	Description of the role of data distributor/owner in data collection and processing	
4	Dates of most recent version of available data	
6	Data refresh/release plans	Documentation of the timing and nature of future updates to the data domains in the data source
7	Supplemental or other data sources available upon request or with additional transformation requirements. Include documentation user interfaces and data collection instruments ('First Instance' data sources)	Examples could include access to narrative text to enable medical chart review or NLP. Access to supply chain data for medical device identification. Access to patient reported outcomes in semi-structured format for outcomes tracking, etc.
Section 2: Data Sources *(repeat for each data source as needed)*		
8	Name of data source and version number or date of extraction	

9	Type of data	e.g. EHR, insurance, registry, patient surveys, spontaneous reports, other (specified)
10	Time frame of data collection	
11	Geographic location	
12	Type of contributing organizations	
13	Level of observation	Patient, encounter, etc
14	Inclusions/exclusion criteria applied at source	
15	Number of unique observations	
16	Number of unique patients (if not at patient level)	
17	Longitudinality of the data	Documentation of the date ranges of availability of different data domains within the data source
18	Basic population summary	e.g. distributions of age, race, ethnicity, sex, geography, insurance coverage, domain-specific high level clinical concepts or exposures
19	Context of data collection	e.g., routine clinical care, trial participation, registry chart abstraction, survey, administrative claims records
20	Surveys implemented	Names and citations for any previously published and validated surveys used
21	Self-reported features	Note any features/variables that are self-reported by patients
22	Unique identifiers/keys	Note any features that on their own or in combination serve as a unique identifier for a patient, device, etc.

		If device unique identifiers are included, note whether lot or serial number are available, UDI/DI/PI available, and how this information is stored
23	Device indication for use	Note whether indication for device use was recorded and format of indication data
24	Permission to use identifiers for further data linkage	Note which identifiers (if any) may be used to link these source data with other data sources and whether permissions are in place to use identifiers for linkage
25	Known reasons for missingness, out of range values, and impact on data completeness	List known reasons for incomplete data collection, proportion of observations impacted, known out of range or erroneous values, and describe subpopulation variance in missingness. E.g., temporal workflow or participation changes
26	Methods for promoting outcome ascertainment	Document proportion of observations without complete follow-up during outcome window and describe any data linkage efforts to improve outcome coverage in the source data
27	Participation rate	Report proportion of anticipated population captured (e.g., proportion of cases or sites included in a registry) or patient response rate
28	Supplemental data generation	If supplemental data are generated as part of the data source collection process specifically for the ETL, careful

		documentation, verification, and accuracy checks and validation should be provided.
29	Informed consent	Document informed consent for secondary use practices and/or obtained waivers of informed consent. Note any limitations of the informed consent process that may introduce bias into the data
30	Privacy preserving data restrictions	Document any filtering, sampling, aggregation, or restrictions used to exclude records in the interest of protecting privacy
31	Requirements for protecting data privacy	Document privacy requirements of the originating data source. E.g., human subjects training for users, storage or security requirements.
32	Relevant regulation/policies	Note any federal, state, local, or source institution regulations/policies that must be adhered to when using these data
33	Evaluation of sample and data collection bias	Document whether any patient subgroups had limited access to care or participation in the settings from which data was collected (e.g., due to insurance status or language barriers). Note observed variation in documentation practices, clinical care practices, study enrollment practices, or patient response rates during data collection
34	Documentation of data collection process deviation	Document deviations from established data capture processes, if present. Document assessments for impact on

		the dataset (e.g., completeness, accuracy, and consistency)
Section 3: Extraction, Transformation, Loading		
35	Documentation of the data model that is being used to represent the data.	If a publicly available data model, reference the version used. If an internally developed model, full documentation of the tables, elements, fields, definitions, and conventions to be used in ETL.
36	Implementation of the data model (ETL process) and documentation of incremental updates	Careful documentation of the implementation of the data model from the data source(s), including any non-standard transformations, conventions, conflicts with the data model definition specifications, and error checking/audit processes. Include data collection and workflows and interfaces to document 'First Instance' data collection.
37	Privacy preservation	Document any privacy preservation processes that are part of the ETL: date shifting, value aggregation, range/transformations, masking, filtering, or sampling
38	Documentation and characterization of controlled vocabularies	Document and characterize controlled vocabularies used by source data and during transformation. This would include descriptions of controlled vocabularies (e.g., ICD-9, ICD-10, CPT/HCPCS, NDC) in use, what versions, and the use of any mapping tools, conventions, or vocabularies used to convert between vocabularies as part of the ETL process (NDC -> RxNorm, ICD9CM -> SNOMED-CT, etc).

39	Data quality evaluations on the transformed (loaded) data	Document assessment of whether standardized data definitions, standardized chart review procedures, and standardized data extraction processes were utilized. If algorithmic or automated tools are used to transform and load data into a data model for reuse, an audit trail with documentation of the algorithms, tools, and error-checks used to perform this task should be documented
40	Summary of missingness	e.g. proportion missing age, race, ethnicity, sex, geography, insurance coverage, domain-specific high level clinical concepts or exposures.
41	Summary of data element duplication or redundancy	Report data element duplication or redundancy.
42	Evaluation of measurement bias	Frequency - Document any variations in missingness across key subpopulations Precision/accuracy - Document any data features that may have varying precision or accuracy in select patient subgroups
43	Data cross-reference validations	Error checking practices through data fields validation, verification or cross-checking against other data fields or data sources
44	Data Storage and use	Document data storage and infrastructure
45	Data dictionary	Attach detailed data dictionary, including for each feature: field name, description, data type, category

		definitions or data standard used, whether mapped to a CDM, representation of missing values, reasonable range of values if applied.
Section 4: Data Linkage		
46	Names and version of linked data sources (refer to section 2)	
47	Purpose of linkage	Describe why data sources were linked
48	Was patient consent required prior to linkage?	Yes/No
49	Was a waiver of informed consent obtained?	Yes/No (if yes, list IRB approving waiver; if no, explain why not)
50	Population overlap	Document how well the data sources overlap and expected proportion of observations that may be linkable
51	Linkage algorithm and process	Provide details for each linkage step, including whether deterministic/probabilistic, features used, thresholds applied, and if/how confidence is reported
52	Privacy protections associated with linkage	Describe privacy protections (one way hashing, tokens). If fully de-identified, provide expert determination review letter.
53	Match rate	Report match rate, overall and by key subgroups
54	Linkage validation methods	Document approach to validating linkages
55	Linkage accuracy	Report linkage accuracy statistics, overall and by key subgroups

56	Linkage dependencies	Were any external (to data ecosystem) tools, environments, and processes required to link data sources? Document all dependencies on data linkage that are not explicitly within the data environment in which the data product resides.
Section 5: Governance		
57	Documentation of the process for user access	Documentation to specify what types of users can reuse the data.
58	Documentation of delay between data collection and downstream operational and research use	Documentation of the approvals, documentations, and regulatory reviews necessary to obtain access.
59	Anticipated data delays	Documentation of delay between data collection and downstream operational and research use
60	Prior use for research	Include citations for research publications using these data
61	Prior use for regulatory decisions	Include citations for regulatory decisions based on these data
62	Permitted uses	Describe permitted use cases
63	Limitations of use	Note specific restrictions on use of entire dataset or components of dataset
64	Requirements for access	Defined necessary training and/or use agreements required to access the data, including contact information for requesting access.
65	Infrastructure requirements	Define any technical requirements for transfers, securely storing, and analyzing date, including data security requirements.

66	Shareholder involvement	Describe any engagement and participation in dataset curation by key shareholder groups (e.g., clinicians, patients)
67	Expertise and training assurance	Documentation of site training, support, and personnel that conduct data collection and transformation
68	Funding disclosure	Disclosure all sources of funding related to data Infrastructure requirements
69	Conflicts of interest	Acknowledge any potential conflicts of interest among key organizations and individuals involved in data Infrastructure requirements
70	Recommended citation for data users	

Note: Table adapted for secondary use medical device data sources from: *Gatto et al. Using real-world data and analytics to support decision making in a global pandemic: Lessons learned. Manuscript under review*