Supplementary file

Table A1. Generic framework for data quality audits in clinical quality registries

Procedure	Overview
Audit of data abstraction	**Definition:** The standardised, structured and systematic program of auditing data abstraction. **Purpose:** To assess data accuracy and completeness, assure data validity and reliability, and guarantee the scientific integrity and academic credibility of derived conclusions. **Implementation:** There are three steps to audits of data abstraction. Step one is to independently replicate the data collection and entry process for a subset of the original registry database over a defined period. This involves abstracting the primary data source for all data fields to produce a replicated dataset. This should be performed by a trained auditor who is blinded to the registry dataset to avoid researcher bias or error due to inexperience. Where it is not possible to replicate all variables, it may be necessary to focus on predetermined variables where levels of agreement or completeness are low. Step two is to analyse the discrepancies between the registry dataset and replicated dataset.[1] If the replicated dataset is assumed to be correct and is considered the gold-standard for data quality, discrepancies represent inaccurate and incomplete data in the registry dataset. In clinical quality registries, there is seldom a gold-standard for data accuracy. Some studies have considered the replicated dataset to be the gold-standard for data quality, however this must be considered in the context of human error and inaccurate or incomplete documentation. Alternatively, if the replicated dataset is not considered the gold-standard for data quality, discrepancies may represent sub-optimal data in either datasets. The original data source may be checked by an independent third party to isolate the sub-optimal data to either dataset, or agreement between the registry dataset and the replicated dataset may be calculated as a proxy indicator for data accuracy. Irrespective of method, overall and item-specific data agreement should be calculated and an attempt should be made to identify the types, causes and frequencies of data discrepancies.[2] Data discrepancies should be resolved by pre-determined arbitration and reconciliation mechanisms.[3] Step three is to calculate the data completeness of the registry dataset and replicated dataset to determine the level of overall, random and systematic incomplete data. It is inevitable that some data imperfections will remain undetected and uncorrected irrespective of quality assurance procedures.[2] **Alternatives:** In assessing data accuracy, there are two, albeit incomplete, alternatives to the time and resource intensive process of auditing data abstraction. One automated and inexpensive alternative is data recoding. This involves reassigning codes to registry data and determining agreement between the recoded and originally coded data.[4] Another alternative is comparing registry datasets against independently collected and validated external datasets, such as local, regional or national administrative datasets.[5] Where possible, the preference is to auto-populate the registry from source data to remove the need for human abstraction and minimise error. Where fiscal constraints are a concern, it may be necessary to prioritise audit of data fields which have previously been reported to be problematic, which are used in the generation of quality indicators to benchmark performance, or which are used to provide risk adjustment.
Implementing method	**Definition:** The method of implementing the data quality audit. **Purpose:** To determine whether auditing data abstraction is performed centrally, remotely or locally. **Implementation:** There are three methods of implementing audits of data abstraction. A central audit of data abstraction is performed by trained auditors from the coordinating centre who visit participating sites to verify and adjudicate data, and audit and review adherence to registry protocols.

	A remote audit of data abstraction may be possible in select circumstances. This resembles a central audit of data abstraction with the exception that the trained auditors perform the audit from the coordinating centre. A local audit of data abstraction is performed on-site by internal data collectors at participating sites. Whilst local audits are potentially less labour-intensive and expensive, they must be preceded by independent validation of self-auditing procedures.
Sampling method	**Definition:** The process of selecting patient records to audit. **Purpose:** To determine whether patient records to audit are selected by single-stage sampling or multi-stage sampling. **Implementation:** Random sampling is recommended to avoid selection bias. There are two methods of random sampling for registries. Single-stage sampling refers to the random selection of patient records from the central database so that each patient record in the registry has an equal probability of being selected.[1] This allows assessment of overall data quality and variations in data quality between participating sites. Multi-stage sampling is a form of stratified sampling that refers to the random selection of participating sites, followed by the random selection of patient records from each selected site.[1] Selected sites are ineligible for reselection until all participating sites have been audited, establishing an audit cycle.[6] This allows limited resources to be concentrated in a select number of participating sites, whilst avoiding bias from selecting sites with specific characteristics.[7] Multi-stage sampling may be modified with convenient selection of participating sites and random selection of patient records from each site, however this may introduce selection bias.
Sample size	**Definition:** The number of patient records to audit. **Purpose:** To determine whether the audit will assess the general picture of data quality or the variations in data quality and its determinants.[1] **Implementation:** There are four methods of determining sample size. Sample size may be determined by absolute number, relative percentage, a combination of the two (e.g. a set percentage of patient records from each participating site, with a minimum number of patient records for participating sites below a predetermined number of patient records and a maximum number of patient records for participating sites above a predetermined number of patient records) or alternative statistical methods (e.g. survey sample size).
Frequency	**Definition:** The regular repetition of formal data quality audits at defined intervals. **Purpose:** To enable the periodic monitoring, maintenance and improvement of data quality. I**mplementation:** Registries are dynamic entities, evolving at a registry, site and patient level. Whilst registry staff are always aware of central changes to procedures, they are not always aware of site-specific changes to procedures or gradual drifts in patient populations or health characteristics.[8] The occasional ad-hoc audit may be appropriate.
Data quality indicators	**Definition:** The determination of the data quality attributes against which data quality will be measured, and the selection of measurable parameters, agreed benchmarks and well-defined standards to quantify the registry's achievement of each data quality attribute. **Purpose:** To measure data quality, benchmark data quality improvement, and allow data comparison. **Implementation:** It is neither possible or practical to set the acceptable data quality standards to 100%. Rather, data quality metrics and acceptable minimum values should be selected based on the intended use of the registry data. Note, despite their ease of interpretation, the use of percentages and proportions in reporting data quality indicators should be considered in the context of their binary nature, inadequate information on the location and spread of distribution, and frequently arbitrary determination.[9] For this reason, if percentages and proportions are used, the quality indicator should be measured

	on a continuous scale, according to several categories, and with acknowledgement of the location and spread of distribution.[9]
Data quality improvement	**Definition:** The standardised process of revising sub-optimal data and improving data quality. **Purpose:** To improve data quality, transparency and accountability **Implementation:** Data quality improvement consists of four steps. Step one is to feedback the results of the data quality audit and recommendations for data quality improvement to data collectors and medical administrators from participating sites.[6] Data quality improvement should be informed by the relationship between presence of sub-optimal data and the processes of data collection, entry and management. Step two is revising incorrect data, incomplete data or incorrect registration in the central registry database.[6] This may involve contacting participating sites to verify data issues, notifying participating sites of data revisions, and tracking corrections to identify recurring issues. Step three is implementing a multifaceted, feasible and timely action plan for data quality improvement of audited sites that fall short of agreed data quality indicators. This should target the local determinants and root causes of sub-optimal data, set goals to be achieved over a designated period, and involve local data collectors.[2] This may involve system and process redesign or re-engineering.[10] Step four is publishing audit reports that document data quality audit findings and quality improvement efforts, and distributing audit reports to stakeholders.[6] The achievement and maintenance of quality improvement implementations should be assessed to determine efficacy.[11] Note, when commenting on differences in data quality over time and across sites, random variation due to chance must be considered.[9]
Qualitative component	**Definition:** The structured interview or survey of healthcare personnel involved in data collection, entry and management.[2] **Purpose:** To achieve insight into the local infrastructure and organisation of data collection and entry, data that are not amenable to audit of data abstraction, and the causes and prevention of sub-optimal data.[2] **Implementation:** Note, as the qualitative component is based on self-report, it may be subject to overestimation. To minimise this potential, participants should be asked to provide examples or evidence supporting claims.
Audit of registry coverage	**Definition:** The standardised, structured and systematic process of reviewing registry recruitment through independent ascertainment of records in all, or a random sample of, participating sites, and matching these records against the registry's data.[3] **Purpose:** To calculate registry coverage and remove duplicate entries. **Implementation:** Where a common pool of identifying variables or unique patient identifiers exist to match information, registry coverage may be estimated by record linkage or cross-referencing to external databases. Duplicate entries may be audited by identification rules for individual patients using combinations of personal characteristics and identification numbers. May include mandatory reporting. May include encouraging private sites to participate.

References

1. Van Der Meulen JHP, Jacob M, Copley L. Assessing the quality of the data in a transplant registry: the European Liver Transplant Registry. *Transplantation* 2003;75(12):2164-67.
2. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9(6):600-11.
3. Couchoud C, Lassalle M, Cornet R, et al. Renal replacement therapy registries-time for a structured data quality evaluation programme. *Nephrology Dialysis Transplantation* 2013;28(9):2215-20. doi: http://dx.doi.org/10.1093/ndt/gft004
4. Mah JK, Warner J, Hall R, et al. Evaluation of neurological patient registries. *Can J Neurol Sci* 2013;40(4 Suppl 2):S60-1.
5. Jacobs ML, Jacobs JP, Franklin RC, et al. Databases for assessing the outcomes of the treatment of patients with congenital and paediatric cardiac disease--the perspective of cardiac surgery. *Cardiol Young* 2008;18 Suppl 2:101-15. doi: https://dx.doi.org/10.1017/S1047951108002813
6. Kuwatsuka Y. Quality control and assurance in hematopoietic stem cell transplantation data registries in Japan and other countries. *International Journal of Hematology* 2016;103(1):20-24. doi: http://dx.doi.org/10.1007/s12185-015-1896-8
7. Tomotaki A, Kumamaru H, Hashimoto H, et al. Evaluating the quality of data from the Japanese National Clinical Database 2011 via a comparison with regional government report data and medical charts. *Surg* 2019;49(1):65-71. doi: https://dx.doi.org/10.1007/s00595-018-1700-5
8. Roger VL. Of the importance of motherhood and apple pie: what big data can learn from small data. *Circulation: Cardiovascular Quality and Outcomes* 2015;8(4):329-31. doi: http://dx.doi.org/10.1161/CIRCOUTCOMES.115.002115
9. Santos M, Eriksson H. Making quality registers supporting improvements: a systematic review of the data visualization in 5 quality registries. *Qual Manag Health Care* 2014;23(2):119-28. doi: https://dx.doi.org/10.1097/QMH.0000000000000021
10. Schwamm L, Reeves MJ, Frankel M. Designing a sustainable national registry for stroke quality improvement. *Am J Prev Med* 2006;31(6 Suppl 2):S251-7. doi: https://dx.doi.org/10.1016/j.amepre.2006.08.013
11. Moussa I, Hermann A, Messenger JC, et al. The NCDR CathPCI Registry: a US national perspective on care and outcomes for percutaneous coronary intervention. *Heart* 2013;99(5):297-303. doi: https://dx.doi.org/10.1136/heartjnl-2012-303379