

Data Quality and Reconstruction in the Age of Artificial Intelligence: Lessons from Cancer Registration and Beyond

Gilbert MacKenzie (Formerly of the University of Limerick, Ireland) &
Nina Briggs (Analytics Engines, Belfast, NI, UK.)

Cancer Registration Data We describe problems that were encountered, when using the 5 national cancer registries to research lung cancer incidence and survival rates in the British Isles between 1993–2018. One registry could not produce annualised incidence and survival rates for the period, while another could not define the data used to create its annualised trend file. The data files arrived in different formats created by different versions of software, some using macros unsupported by the Excel clone in use. This, despite all 5 being members of the United Kingdom and Ireland Association of Cancer Registries (UKIACR), which facilitates research. The lack of common standards (harmonisation) in data processing structures and sharing delayed the research. This experience raises questions about customer service, staff competence and professional focus on the basics of Cancer Registration which may impede maintenance of up-to-date National Statistics. This is only one example in Cancer Registration, but the problems are more wide-ranging and the implications for AI obvious.

AI: Data quality and the limits of reconstruction. Artificial intelligence can compensate for some failures of standardisation, but it cannot necessarily resolve uncertainty about what data mean. AI can assist with extracting information from heterogeneous files, free text and complex decision rules, but these capabilities should not be confused with clinical understanding. Models may lack tacit clinical knowledge, relevant information, or the immediate feedback available through clinical interaction. An important distinction therefore exists between data which are difficult to process computationally and data whose meaning is genuinely uncertain. AI-assisted reconstruction may involve extraction, inference or imputation, each with different evidential implications. Treating these as equivalent to directly observed information may improve apparent completeness while reducing transparency and obscuring provenance. Missingness may itself contain information about the healthcare system. Patterns in what is recorded, for whom and when may reflect healthcare utilisation and observation processes; such informative presence can bias analyses of routinely collected electronic health record data. Historical funding, access, service provision and research priorities also influence which populations generate abundant data. Models may therefore reproduce patterns generated by healthcare systems rather than underlying clinical phenomena. For example, a widely used population-health algorithm exhibited substantial racial bias because healthcare expenditure was used as a proxy for health need. These problems are particularly important for smaller or historically under-represented populations, where limited observations constrain estimation and validation. They also become more complex when predictions influence healthcare provision: the model then becomes part of the future data-generating process. AI should therefore expose complexity rather than conceal it: preserving provenance, distinguishing observed from inferred information, retaining meaningful missingness, representing uncertainty and allowing abstention where appropriate.