Royal Statistical Society (RSS) Data Ethics Special Interest Group (SIG) – Written evidence (COV0048)

Submitted by Thomas King, Secretary, RSS Data Ethics SIG

Among the member activities of the Royal Statistical Society (RSS) is a Special Interest Group (SIG) on Data Ethics. The SIG was formed by RSS Council in June 2018. In January 2019 the SIG published a working paper on the data ethics landscape. Our work is summarised below.

Data Ethics: a statistical perspective

Data now originates as a byproduct of transactions, administrative systems, and new technology. Ethical issues surround the settings of origin, impacts of systems, and the people involved. Other approaches, specifically bioethics and digital ethics, miss the statistical issues around systems:

- Bioethics is concerned for individuals, so even the justice principle is natural, not social.
- Digital ethics attends to technology, privacy and data stored or transmitted, not statistics.

Data suitable for statistical analysis arises for a particular purpose, coded to relevant conventions. Ethical choices ought to be made in its reuse: repurposing; reanalysis; and recombining (linkage).[1]

- Purpose classes include: transactions; accounting; audit; evaluation; research; marketing.
- Reanalysis is use under original terms by new groups, with a suitable governance process.
- Linkage means augmenting datasets statistically[2], for use in further analysis types above.

In general use involves choices about statistical assumptions used in models. These choices are about probabilistic structures and categories which read across to objects and down to individuals.

Three organising values for statistical work of this kind arise: justice, respect and solidarity.

- Justice concerns the structural patterns in society which are represented in any statistical analysis. Ranging from sample coverage and sparseness of data to model specification. This includes assessments of discriminatory outcomes for people in protected categories.
- Respect concerns how individuals are represented in data. This includes attention to what they consent to, in terms of inclusion, and the categories describing their circumstances. It extends to attention for pejorative conclusions about social groups (through their status).

- Solidarity represents the combined concern that groups have for the use of data about them, the impact its conclusions have, and how they are represented. These groups may include similar people, those in similar circumstances or location, or people in the future.[3]

Communicating effectively with the public about probability and structural patterns relating to abstract models of society is not easy. Public benefit is determined by paternalistic professional and regulator authority. Establishing a social licence for use of data is considered good practice.[4] Trustworthy communication relies on the idea of intelligent transparency[5]: Information is relevant, interpretable, accessible and assessable. News media have a role, as do standards of evidence.[6] New technology presents a particular challenge. Contexts previously inscrutable to analysis now serve to generate extensive data. This challenges the contextual integrity of the social.[7] Systems for travel are helpful planning journeys and monitoring delays, but commuting may reveal a home location. Social networks are typically intended to be shared at limited remove 'friends of friends'. Data has great reuse potential, especially linked together, which may undermine personal privacy.

Ethical governance of data promotes agreed use as a club good (data trusts) or public good. The prevailing practice is much more about established uses and internal use of data (without adequate safeguards). Linking data, even in clear pursuit of public interest, is restricted. Stewardship of data by disinterested oversight boards, with attention to all stakeholders is needed[8] but the role models to suit a virtue ethics[9] are lacking. Capability throughout the data ecosystem should be developed.

Public Communication

Each of the three issues below affects the public, choices being made to use data ought to reflect values of justice, respect and solidarity. Distinct data use purposes ought to be evaluated for the structure imposed on society. Public communication, and accountability, has been a sustained weakness of the work using data in the current crisis: trustworthiness ought to be a universal aim.

Ethnicity

As in this case, infection models typically include some demographic information, about age / sex structure of a population. Ethnicity was not explicitly mentioned in SAGE minutes until April, although some research was in progress by that point. Health outcomes of every kind have many social causes and consequences, but persistent social inequalities show justice is not achieved.[10] Ethnicity is about identity and cultural heritage, complex issues not easily specified in statistical categories large but consistent enough for analysis. ONS took the unprecedented step of using the 2011 census to look up characteristics such as ethnicity and religion of registered deaths.[11] This uses two steps of matching the census data to the demographic population spine from NHS Digital using statistical research provisions, then matching the deceased individuals to

the population spine and reading across. While GDPR only applies to living natural persons, this is at odds with a 100 year confidentiality standard. Inquiries (of NSDEC) about the ethical review are unanswered.

Local Surveillance

Data about cases and testing have several uses, one of which is local surveillance. This is distinct from diagnostics, individual treatment, contact tracing, backward analysis, test provision / evaluation and (health) research. Each of these purposes is different, and ethical consideration ought to be made of the propriety of each data use. Where the justification for use of the data is a public benefit, a clear account of the evaluation of the benefits ought to be available, including where this means developing an evidence base. Local surveillance ought to involve a calibrated and coherent analysis of data from several sources, comparing observed to expected outcomes, including impersonal data from wastewater. Simply handing the personal details of every person tested to local authorities will not achieve this: such demands take social licence for granted. Interventions will need to be escalated where clusters spread to outbreaks of renewed community transmission. Coherent approaches to modelling, acute at times of outbreak, may be most pragmatic at region level.[12] Communication about data use ought to be intelligently transparent.

Modelling

Models, mathematical and statistical, have been used extensively. It is still very common to hear that there is no ethical content in mathematics[13], whether theorems are canonical or the ethics only appears in the applications. Standard professional issues of creating models without domain expertise, or publishing them without suitable expert review were raised early on. The question of which models are used in urgently commissioned expert analysis is also acknowledged. But how expediency has forced choices, and potential effects, ought to be clearly communicated to all who may use a model. Where outcomes are restricted, such as managing acute hospital admissions, the effect of excluding other outcomes should be investigated, as what is not modelled is not seen. Similarly, data must depend on structures and categories available, but those not used, despite substantive interest, for want of data, should be considered for further study. Metapopulations[14] of structures beyond schools and hospitals apply, as do social circumstances of multigenerational households. Similarly modelling disease states of recovery and mild / moderate case severity.

References

1. https://link.springer.com/chapter/10.1007/978-3-319-28422-4_2
2. https://academic.oup.com/jpubhealth/article/40/1/191/3091693
3. https://www.cambridge.org/core/books/solidarity-in-biomedicine-and-beyond/067DC974D204F6EDE679816213433456
4. https://jme.bmj.com/content/41/5/404
5. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

6. https://hdsr.mitpress.mit.edu/pub/56lnenzj/release/1
7. https://www.sup.org/books/title/?id=8862
8. https://humgenomics.biomedcentral.com/articles/10.1186/s40246-018-0154-6
9. https://global.oup.com/academic/product/technology-and-the-virtues-9780190498511
10. https://www.health.org.uk/publications/reports/the-marmot-review-10-years-on
11. https://www.ons.gov.uk/releases/coronaviruscovid19relatedmortalitybyreligionethnicityanddisabilityenglandandwales2march2020to15may2020
12. https://theconversation.com/coronavirus-why-we-need-local-models-to-successfully-exit-lockdown-138358
13. https://ethics.maths.cam.ac.uk/assets/dp/18_1.pdf
14. https://www.sciencedirect.com/science/article/pii/S175543651400036X

15 July 2020