5 Safes of secure access to confidential data

Felix Ritchie
Bristol Economic Analysis
University of the West of England

Safe outputs:

Principles, rules, risks and expectations

Administrative Data Research Network

UK Data Service

HM Revenue & Customs

Justice

Office for National Statistics

The Health Foundation

How should we set speed limits?

- Consider setting 30mph as the speed limit for all urban roads
  - What are the advantages and disadvantages?

30mph limit on all urban roads

- Advantages
  - Clear and memorable
  - Simple
  - Easy to enforce
  - Consistent in a range of settings
  - Fair

- Disadvantages
  - Not appropriate for every context
    - Too low in some cases => **inefficient**
    - Too high in some cases => **unsafe**
  - May not have credibility

Plan B: a rule-of-thumb

- Set 30mph as the default, but letting urban planners set different limits depending on road conditions
  - Advantages? Disadvantages?

30mph limit by default but not always

- Advantages
  - Responsive to circumstances
  - More efficient
  - Safer
  - Likely to have more public acceptability
- Disadvantages
  - Need to notify
  - More complicated to explain
  - More complicated to enforce
  - Demands more engagement from everyone
- The ‘principled’ approach is safer and more efficient
  - but more complex

So which is better?

- Strict universal rule:
  - simpler and cheaper
  - doesn’t need skill in assessing road risk
  - no room for argument or ambiguity

- Broad principles and specific solutions:
  - greater efficiency in road use
  - greater safety
  - forces evaluation of risk
  - forces subjective decisions about resource allocation
  - principles easily extensible

From analogy to practice

- Rule:
  - ‘A total based on 5 respondents is okay for release’
- Principle:
  - ‘A total based on 5 respondents is probably okay’
  - but sometimes that may be too low, and sometimes too high
  - use expertise to confirm what is appropriate in this specific case
  - key advantage:
    - make defaults relatively strict (eg limit of 10) => greater safety
    - change depending on circumstance => greater efficiency
- Two problems with the principles-based approach:
  - expertise
  - number of outputs to be checked

Problem 1: expertise in output checking

- Both parties need to have some understanding of how and why rules-of-thumb are adapted

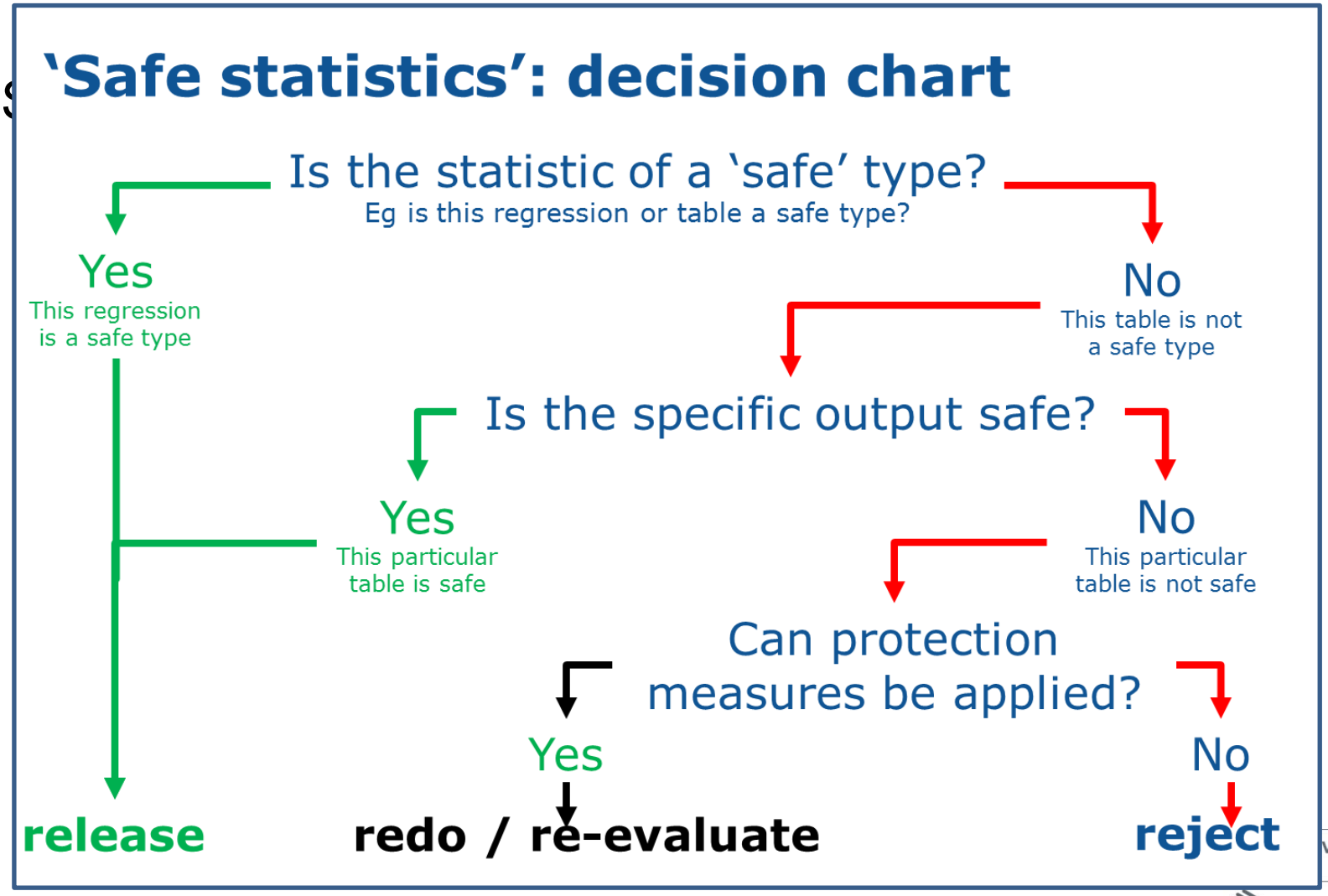
  $\Rightarrow$ need for training of both

- Disadvantage:
  - cost of training
  - feasibility

- Mitigating factors:
  - many existing solutions to build on
  - approach is positively welcomed by researchers
  - opportunity to develop relationship with researchers

Problem 2: number of outputs

- S



Uncertain perspectives

- Explicitly now a ‘balance of risks’ model
  - ‘safe’ is low risk, not zero risk
  - resources matter – concentrate on high-risk outcomes
- Data owners don’t like this, preferring certainty of rules
  - but no-one actually applies a strict rules model in practice…
  - being explicit stimulates consideration of training
- Researchers do
  - it reflects their understanding of risk
  - it reflects statistical practice
- Again, relationship with users is key
  - evidence shows researchers happy to self-police

Where are we now?

- Principles-based output checking and ‘safe statistics’:
  - broadly accepted as best practice for secure facilities
  - lots of practical experience in doing
  - lots of practical experience in training
  - developing suite of training materials

- Output-checking aligns with good statistical practice
  - especially when principles-based
  - aligns with researcher experience and understanding
  - training provides another opportunity to build relationship

References

- Principles-based output SDC
  - Brandt et al (2010), *Guidelines for the checking of output based on microdata research*
  - Ritchie F. and Elliot M. (2015) “Principles- versus rules-based output statistical disclosure control in remote access environments”
  - Ritchie F. and Welpton R. (2015) “Operationalising Principles-based Output SDC” – *draft, comments very welcome*

- Safe statistics
  - Ritchie F. (2008) “Disclosure detection in research environments in practice”
  - Ritchie F. (2014) “Operationalising safe statistics: the case of linear regression”