



KEELE
UNIVERSITY



Evidence-Based Software Engineering

Barbara Kitchenham

Tore Dybå (SINTEF)

Magne Jørgensen (Simula Laboratory)

Agenda

- The evidence-based paradigm
- Evidence-Based Software Engineering (EBSE)
 - Goals
 - Procedures
- Comparison with evidence-based medicine
- Conclusions

The Evidence-Based Paradigm

- Evidence-based medicine has changed research practices
 - Medical researchers found
 - Failure to organise existing medical research cost lives
 - Clinical judgement of experts worse than systematic reviews
- Evidence-based paradigm adopted by many other disciplines providing service to public
 - Social policy
 - Education
 - Psychiatry

Impact of EBM

- 1992
 - 1 publication on EBM
- 1998
 - 1000 publications
 - 6 journals
 - Specialising in evidence-based medicine
- Criticisms
 - Research is fallible
 - Relies on generalisations that may not hold
 - Often insufficient to determine appropriate practice
 - Software issue –speed of technology change

Evidence-Based Software Engineering (EBSE)

- Research question
 - Is evidence-based paradigm feasible for Software Engineering?
 - “Everyone else is doing it”
 - Not a valid argument
- Methodology
 - Analogy-based Comparison
 - Evidence-based paradigm in medicine v software engineering

Goal of EBSE

- EBM: Integration of best research evidence with clinical expertise and patient values
- EBSE: Adapted from Evidence-Based Medicine
 - To provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision making process regarding the development and maintenance of software
- Might provide
 - Common goals for research groups
 - Help for practitioners adopting new technologies
 - Means to improve dependability
 - Increase acceptability of software-intensive systems
 - Input to certification process

Practicing EBM & EBSE

- Sets requirements on practitioners and researchers
- Practitioners
 - Need to track down & use best evidence in context
- Researchers need to provide best evidence

What is Evidence?

- Systematic reviews
 - Methodologically rigorous synthesis of all available research relevant to a specific research question
 - Not ad hoc literature reviews
- Best systematic reviews based on Randomised Controlled Trials (RCTs)
 - Not laboratory experiments
 - Trials of real treatments on real patients in a clinical setting
 - Most (perhaps all) SE experiments are laboratory experiments

Integrating evidence

- Medical researchers & practitioners construct practitioner-oriented guidelines
 - Assess the evidence
 - Determine strength of evidence (type of study)
 - Size of effects (practical not just statistical)
 - Relevance (appropriateness of outcome measures)
 - Assess applicability to other settings
 - Summarise benefits & harms
 - Present the evidence to stakeholders
 - Balance sheet

Medical Infrastructure – 1/2

- Major databases of abstracts & articles
 - Medline (4600 biomedical journals)
- 6 evidence-based journals specialising in systematic reviews
- Cochrane collaboration
 - Database of systematic reviews (RCT-based)
 - <http://www.cochrane.org>
- Campbell Collaboration for social policy

Medical Infrastructure – 2/2

- Standards to encourage experimental rigour & improve accumulation of evidence
 - Individual empirical studies
 - Based on agreed experimental guidelines
 - Reporting standards
 - Including structured abstracts
 - Systematic Reviews
 - Guidelines for assembling, collating and reporting evidence
 - Evidence-based guidelines for practitioners
 - Developed by mixed panels
 - Practitioners, Researchers, Methodologists, Patients

Software Engineering

- No comparable research infrastructure
- No agreed standards for empirical studies
 - A proposal for formal experiments and surveys
 - Nothing for qualitative or observational studies
 - No agreed standards for systematic review
 - Kitchenham Technical report adopted by IST
- Few software engineering guidelines based on empirical evidence
 - CMM has been back-validated but wasn't itself based on evidence
 - Contrast with guidelines for Web apps

Scientific Issues- 1/2

- The skill factor
 - SE methods usually require a trained individual
 - Can't blind subject to treatment
 - Can't control for experimenter and subject expectations
- Need to improve protocols
 - Use blinding whenever possible
 - Replicate experiments
 - But not too closely
- Need to qualify our experiments
 - Strength of evidence is less for laboratory experiments

Scientific Issues –2/2

- The lifecycle issue
 - Techniques interact with other techniques over a long period of time
 - Difficult to determine causal links between techniques and outcomes
 - Intermediate outputs of a specific task may not be meaningful to practitioners
 - Improved reliability can't be demonstrated in a design document

Addressing Lifecycle issues

- Experiments on techniques in isolation
 - Still have problem that outcomes are not practitioner-relevant
- Large-scale empirical studies
 - Hard to generalise because context is critical
 - Quasi-experiments similar to experiments but without randomisation
 - Need arguments to justify causality
 - Benchmarks based on data from a variety of projects
 - Difficulty with representativeness

Conclusion

- ESBE lacks the infrastructure required to support evidence-based paradigm
 - Would need financial support to put in place appropriate infrastructure
- Scientific problems more intractable
 - Need to develop appropriate protocols for SE studies
- Some aspects of EBSE easy to adopt
 - Systematic review
 - Requirement of every PhD student
 - Procedures can be adopted from medicine
 - Structured abstracts
- EBSE needs to be tested on real problems

Systematic Reviews - 1/2

- A systematic (literature) review is
 - An overview of research studies that uses explicit and reproducible methods
- Systematic reviews aim to synthesise existing research
 - Fairly (without bias)
 - Rigorously (according to a defined procedure)
 - Openly (ensuring that the review procedure is visible to other researchers)

Advantages

- Provide information about effects of a phenomenon across wide range of settings
 - Essential for SE where we have sampling problems
 - Consistent results provide evidence that phenomena are
 - Robust
 - Transferable
 - Inconsistent results
 - Allow sources of variation to be studied
- Meta-analysis possible for quantitative studies

Anticipated Benefits

- Create a firm foundation for future research
 - Position your own research in the context of existing research
- Close areas where no further research is necessary
- Uncover areas where research is necessary
- Help the development of new theories
 - Identify common underlying trends
 - Identify explanations for conflicting results
- Should be a standard research methodology

Disadvantages

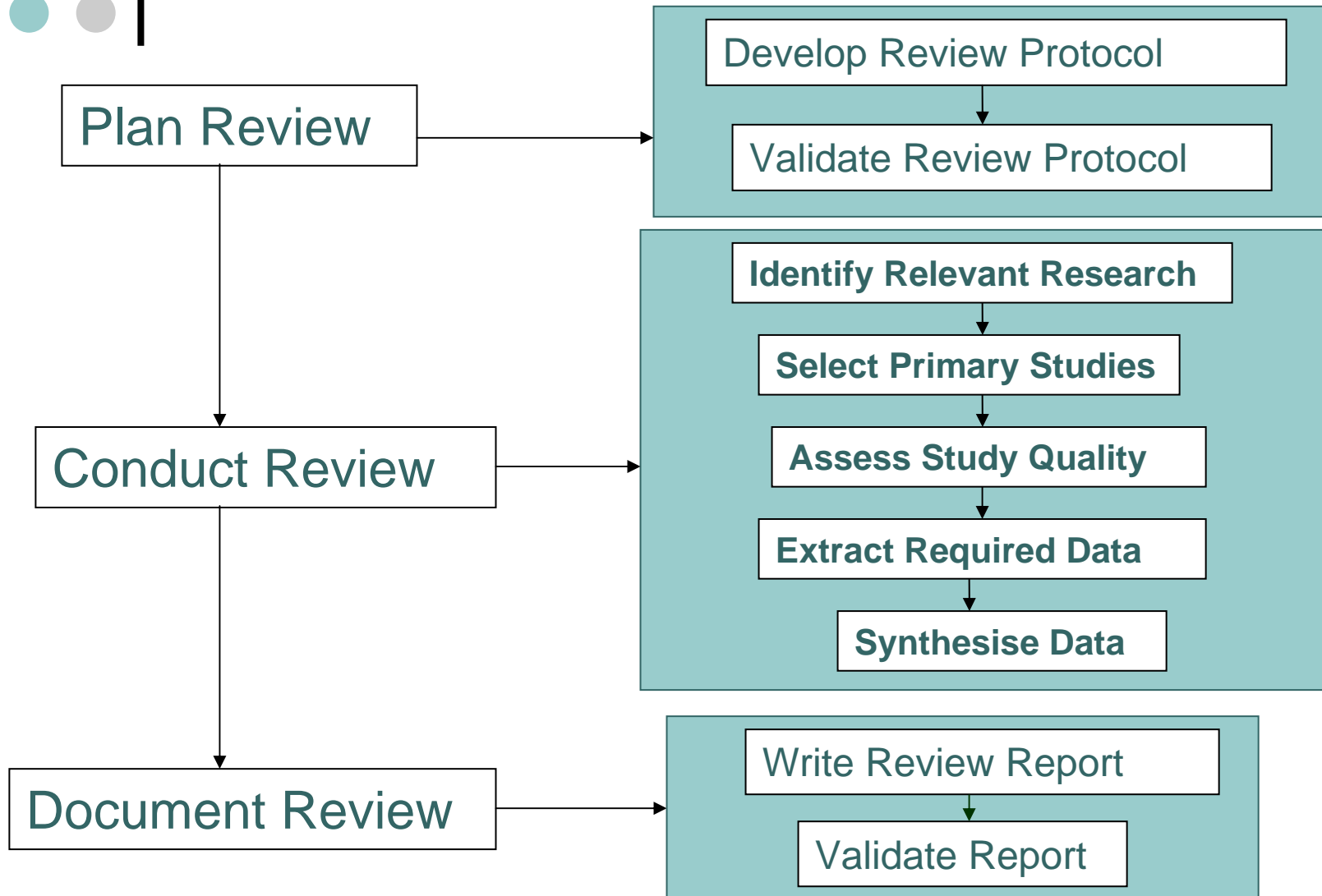
- Require more effort than informal reviews
- Difficult for lone researchers
 - Standards require two researchers
 - Minimising individual bias
- Incompatible with requirements for short papers

Value of Systematic Reviews

- Can contradict “common knowledge”
 - Jørgensen and Moløkken reviewed surveys of project overruns
 - Standish CHAOS report is out of step with other research
 - May have used inappropriate methodology
 - Jørgensen reviewed evidence about expert opinion estimates
 - No consistent support for view that models are better than human estimators



Systematic Review Process



References

Australian National Health and Medical Research Council. How to review the evidence: systematic identification and review of the scientific literature, 2000. ISBN 186-4960329 .

Australian National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. February 2000, ISBN 0 642 43295 2.

Cochrane Collaboration. Cochrane Reviewers' Handbook. Version 4.2.1. December 2003.

Glass, R.L., Vessey, I., Ramesh, V. Research in software engineering: an analysis of the literature. IST 44, 2002, pp491-506

Magne Jørgensen and Kjetil Moløkken. How large are Software Cost Overruns? Critical Comments on the Standish Group's CHAOS Reports, http://www.simula.no/publication_one.php?publication_id=711, 2004.

Magne Jørgensen. A Review of Studies on Expert Estimation of Software Development Effort. Journal Systems and Software, Vol 70, Issues 1-2, 2004, pp 37-60.

References

Khan, Khalid, S., ter Riet, Gerben., Glanville, Julia., Sowden, Amanda, J. and Kleijnen, Jo. (eds) Undertaking Systematic Review of Research on Effectiveness. CRD's Guidance for those Carrying Out or Commissioning Reviews. CRD Report Number 4 (2nd Edition), NHS Centre for Reviews and Dissemination, University of York, IBSN 1 900640 20 1, March 2001.

Kitchenham, Barbara. Procedures for Performing Systematic Reviews, Joint Technical Rreport, Keele University TR/SE-0401 and NICTA 0400011T.1, July 2004. (There s now a revised version)

Pai, Madhukar, McCulloch, Michael, Gorman, Jennifer D., Pai, Nitika, Enanoria, Wayne, Kennedy, Gail, Tharyan, Prathap, Colford, John M. Jr. Systematic reviews and meta-analysis: An illustrated, step-by-step guide. The National medical Journal of India, 17(2) 2004, pp 86-95.

References

- Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W., and Haynes, R.B. *Evidence-Based Medicine: How to Practice and Teach EBM*, Second Edition, Churchill Livingstone: Edinburgh, 2000.
- Sanjay J. Koyani, Robert W. Balley, Janke R. Nall.. Research-based Web Design & Usability Guides, National Cancer Institute. 2003, <http://usability.gov/pdfs/guidelines.html>.