Directrices para la realización de revisiones sistemáticas de la literatura en ingeniería de software

Versión 2.3

Informe técnico EBSE
2007-01

Grupo de Ingeniería de Software
Escuela de Informática y Matemáticas de la
Universidad de Keele
Keele, Staffs
ST5 5BG,
Reino Unido

y

Departamento de Ciencias
Informáticas Universidad de
Durham
Durham,
Reino
Unido

9 de julio de 2007

© Kitchenham, 2007

0. Sección de control de documentos

0.1 Contenido

0.2 Control de la versión del documento

Estado del documento	Número de versión	Fecha	Cambios de la versión anterior
Borrador	0.1	1º de abril de 2004	Ninguno
Publicado	1.0	29 de junio de 2004	Corrección de errores tipográficos Debate adicional sobre problemas de evaluación de las pruebas. Se añadió la sección 7 "Observaciones finales".
Revisión	1.1	17 de agosto 2005	Correcciones de errores tipográficos.
Revisión Mayor	1.9	25 de octubre 2005	Cambiado el título, añadido el SMC como revisor, añadidas varias secciones nuevas, la finalización de las principales revisiones debería ser la versión 2.0 Cambios resumidos a continuación: Añadida la sección 2 - a pero EBSE en Contexto Amplió la presentación de informes de los procesos de examen Se añadieron secciones sobre la cartografía sistemática y los exámenes terciarios en la sección 4 Actualizado la sección de informes de la revisión Se añadieron dos secciones finales, Estudios de Cartografía Sistemática y Revisiones Terciarias

Otras revisiones importantes	2.0	17 de marzo de 2007	Revisó la sección sobre la jerarquía de los estudios para ser coherente con los puntos de vista de las ciencias sociales. Eliminó algunas discusiones generales que no estaban bien enfocadas en la construcción de las directrices. Revisó la sección de listas de control de calidad Se añadieron las lecciones aprendidas de los artículos del SE. Se eliminó la sección final de las revisiones del Terciario (parecía innecesario)
Revisiones menores después de la interna...	2.1	27 de marzo de 2007	Corrección de errores tipográficos Inclusión de un glosario Inclusión de directrices

revisar			proceso de construcción Reestructuraciones menores - Los exámenes cartográficos y los exámenes terciarios pasaron a la sección 3 para no interferir con el flujo de las directrices.
Otras revisiones menores	2.2	4 de abril de 2007	Tipografías y correcciones gramaticales. Un párrafo sobre cómo leer las directrices incluidas en la Introducción.
Revisiones después de las revisiones externas revisar	2.3	20 de julio	Las enmiendas después del examen externo, incluida la introducción de más ejemplos.

0.3 Equipo de desarrollo de documentos

Este documento fue revisado por los miembros del Proyecto de Ingeniería de Software Basado en la Evidencia (EBSE) (EP/CS51839/X), que fue financiado por el Consejo de Investigación de Ciencias Físicas y Económicas del Reino Unido.

Nombre	Afiliación	Papel
Barbara Kitchenham	Universidad de Keele, Reino Unido	Autor principal
Stuart Charters	Universidad de Lincoln, NZ	Segundo autor
David Budgen	Universidad de Durham, Reino Unido	Revisor interno del EBSE
Pearl Brereton	Universidad de Keele, Reino Unido	EBSE Interno Revisor
Mark Turner	Universidad de Keele, Reino Unido	Revisor interno del EBSE
Steve Linkman	Universidad de Keele, Reino Unido	Revisor interno del EBSE
Magne Jørgensen	Investigación de Simula Laboratorio, Noruega	Revisor externo
Emilia Mendes	Universidad de Auckland, Nueva Zelandia	Revisor externo
Giuseppe Visaggio	Universidad de Bari, Italia	Revisor externo

0.4 Resumen ejecutivo

El objetivo de este informe es proponer directrices exhaustivas para la revisión sistemática de la literatura apropiada para los investigadores de ingeniería de software, incluidos los estudiantes de doctorado. Una revisión sistemática de la literatura es un medio de evaluar e interpretar toda la investigación disponible relevante para una cuestión de investigación, un área temática o un fenómeno de interés particular. Las revisiones sistemáticas tienen por objeto presentar una evaluación justa de un tema de investigación mediante el uso de una metodología fiable, rigurosa y auditable.

Las directrices que se presentan en este informe se derivan de tres directrices existentes utilizadas por los investigadores médicos, dos libros producidos por investigadores con formación en ciencias sociales y debates con investigadores de otras disciplinas que participan en la práctica basada en la evidencia. Las directrices se han adaptado para reflejar los problemas específicos de la investigación en ingeniería de programas informáticos.

Las directrices abarcan tres fases de un examen sistemático de la literatura: la planificación del examen, la realización del examen y la presentación de informes sobre el examen. Proporcionan una descripción de nivel relativamente alto. No tienen en cuenta el impacto de las preguntas de la investigación en los procedimientos de revisión, ni especifican en detalle los mecanismos necesarios para realizar un metaanálisis.

0.5 Glosario

Meta-análisis. Una forma de estudio secundario en la que la síntesis de la investigación se basa en métodos estadísticos cuantitativos.

Estudio primario. (En el contexto de las pruebas) Un estudio empírico que investiga una cuestión de investigación específica.

Estudio secundario. Estudio que revisa todos los estudios primarios relacionados con una cuestión de investigación específica con el objetivo de integrar/sintetizar la evidencia relacionada con una cuestión de investigación específica.

Análisis de sensibilidad. Procedimiento de análisis destinado a evaluar si los resultados de un examen sistemático de la literatura o de un metaanálisis están indebidamente influidos por un pequeño número de estudios. Los métodos de análisis de sensibilidad consisten en evaluar el impacto de los estudios de gran influencia (por ejemplo, estudios de gran tamaño o estudios con resultados atípicos) y garantizar que los resultados generales de una bibliografía sistemática sigan siendo los mismos si se omiten del análisis los estudios de baja calidad (o de alta calidad) o se analizan por separado.

Revisión sistemática de *la literatura* (también denominada revisión sistemática). Una forma de estudio secundario que utiliza una metodología bien definida para identificar, analizar e interpretar todas las pruebas disponibles relacionadas con una cuestión de investigación específica de manera imparcial y (hasta cierto punto) repetible.

Protocolo de *revisión sistemática*. Un plan que describe la realización de una propuesta de revisión sistemática de la literatura.

Estudio cartográfico sistemático (también llamado estudio de alcance). Un examen amplio de los estudios primarios en un área temática específica que tiene por objeto identificar las pruebas disponibles sobre el tema.

Estudio terciario (también llamado revisión terciaria). Una revisión de estudios secundarios relacionados con la misma cuestión de investigación.

1. Introducción

En este documento se presentan directrices generales para la realización de exámenes sistemáticos. El objetivo de este documento es presentar a la comunidad de ingenieros de software la metodología para realizar exámenes rigurosos de las pruebas empíricas actuales. Está dirigido principalmente a los investigadores de ingeniería de software, incluidos los estudiantes de doctorado. No cubre los detalles del meta-análisis (un procedimiento estadístico para sintetizar los resultados cuantitativos de diferentes estudios), ni discute las implicaciones que los diferentes tipos de preguntas de revisión sistemática tienen en los procedimientos de investigación.

El impulso original para emplear la práctica de la revisión sistemática de la literatura fue apoyar la medicina basada en la evidencia, y muchas directrices reflejan este punto de vista. Este documento intenta construir pautas para realizar revisiones sistemáticas de la literatura que sean apropiadas para las necesidades de los investigadores de ingeniería de software. Discute una serie de cuestiones en las que la investigación en ingeniería de software difiere de la investigación médica. En particular, la investigación en ingeniería de programas informáticos tiene relativamente poca investigación empírica en comparación con el ámbito médico; los métodos de investigación utilizados por los ingenieros de programas informáticos no son tan rigurosos en general como los utilizados por los investigadores médicos; y muchos datos empíricos de la ingeniería de programas informáticos están protegidos por derechos de propiedad intelectual.

1.1 Material de base utilizado en la construcción de las directrices

El documento se basa en un examen de tres directrices existentes para las revisiones sistemáticas, las experiencias del proyecto de ingeniería de software basado en la evidencia de la Universidad de Keele y la Universidad de Durham, las reuniones con expertos en diversas disciplinas interesados en la práctica basada en la evidencia y los libros de texto que describen los principios de la revisión sistemática:

- El Manual del Revisor Cochrane [7] y el Glosario [8].
- Directrices preparadas por el Consejo Nacional de Salud e Investigación Médica de Australia [1] y [2].
- Directrices del Centro de Examen y Difusión (CRD) para quienes realizan o encargan exámenes [19].
- Revisiones sistemáticas en las Ciencias Sociales: Una guía práctica, Mark Petticrew y Helen Roberts [25]
- Realizando revisiones de literatura de investigación. From the Internet to Paper, ^{2a} edición, Arlene Fink [11].
- Diversos artículos y textos que describen los procedimientos para las revisiones de la literatura en medicina y ciencias sociales ([20], [13] y [24]).
- Reuniones con diversos expertos y centros del ámbito, entre ellos, el Centro de Información y Coordinación de Pruebas para la Política y la Práctica (EPPI Centre http://eppi.ioe.ac.uk/cms/) Unidad de Investigación en Ciencias Sociales Instituto de Educación, Universidad de Londres; CRD Universidad de York, Mark Petticrew, Universidad de Glasgow; Andrew Booth, Universidad de Sheffield
- Experiencias del Proyecto de Ingeniería de Software Basado en la Evidencia

en la Universidad de Keele y la Universidad de Durham.

En particular, este documento debe mucho a las Directrices del CRD.

1.2 El proceso de construcción de la guía

El proceso de construcción utilizado para las directrices fue:

- Las directrices fueron originalmente producidas por una sola persona (Kitchenham).
- Luego fueron actualizadas por dos personas (Charters y Kitchenham).
- Fueron revisados por miembros del proyecto de Ingeniería de Software Basada en la Evidencia (Brereton, Budgen, Linkman y Turner).
- Tras la corrección, las directrices se distribuyeron a expertos externos para su examen independiente.
- Las directrices se volvieron a modificar tras el examen de los expertos externos.

1.3 La estructura de las directrices

La estructura de las directrices es la siguiente:

- La sección 2 ofrece una introducción a los exámenes sistemáticos.
- En la sección 3 se explica por qué la metodología SLR de las ciencias sociales es apropiada en el contexto de la investigación de la ingeniería de software.
- En la sección 4 se especifican las etapas de un examen sistemático.
- La sección 5 trata de las etapas de planificación de un examen sistemático.
- En la sección 6 se analizan las etapas que intervienen en la realización de un examen sistemático.
- La sección 7 trata sobre la presentación de informes de un examen sistemático.
- La sección 8 trata de los estudios de cartografía sistemática.

A lo largo de las directrices hemos incorporado ejemplos tomados de dos revisiones sistemáticas de la literatura recientemente publicadas [21] y [17]. Kichenham y otros [21] abordaron la cuestión de si era posible utilizar conjuntos de datos de evaluación comparativa entre empresas para producir modelos de estimación adecuados para su uso en una empresa comercial. Jørgensen [17] investigaron el uso del juicio de los expertos, los modelos formales y las combinaciones de los dos enfoques al estimar el esfuerzo de desarrollo de software. Además, en el apéndice 2 figura una lista de las revisiones sistemáticas de la literatura publicada que los autores del presente informe han evaluado como de alta calidad. Estas SLRs fueron identificadas y evaluadas como parte de una revisión bibliográfica sistemática de SLRs recientes de ingeniería de software. El protocolo de la revisión está documentado en el Apéndice 3.

1.4 Cómo utilizar las pautas

Estas directrices están dirigidas a investigadores de ingeniería de software, estudiantes de doctorado y profesionales que son nuevos en el concepto de realizar revisiones sistemáticas de la literatura. Los lectores que no estén seguros de lo que es una revisión sistemática de la literatura deben empezar por leer la Sección 2.

Los lectores que entiendan los principios de una revisión sistemática de la literatura pueden saltar a la Sección 4 para obtener una visión general del proceso de revisión sistemática de la literatura. A continuación, deben concentrarse en las secciones 5, 6 y 7, que describen en detalle cómo realizar cada fase de revisión. Las secciones 3 y 8 proporcionan información auxiliar que puede ser omitida en la primera lectura.

Los lectores que tengan más experiencia en la realización de exámenes sistemáticos

podrán encontrar que la lista de tareas de la sección 4, las listas de control de calidad de los cuadros 5 y 6 y la estructura de presentación de informes presentada en el cuadro 7 son suficientes para sus necesidades.

Es poco probable que los lectores con preguntas metodológicas detalladas encuentren respuestas en este documento. Pueden encontrar útiles algunas de las referencias.

2. Revisiones sistemáticas de la literatura

Una revisión sistemática de la literatura (a menudo denominada revisión sistemática) es un medio de identificar, evaluar e interpretar todas las investigaciones disponibles que sean pertinentes a una cuestión de investigación particular, o a un área temática, o a un fenómeno de interés. Los estudios individuales que contribuyen a un examen sistemático se denominan estudios primarios; un examen sistemático es una forma de estudio secundario.

2.1 Razones para realizar revisiones sistemáticas de la literatura

Hay muchas razones para realizar una revisión sistemática de la literatura. Las razones más comunes son:

- Resumir las pruebas existentes relativas a un tratamiento o tecnología, por ejemplo, resumir las pruebas empíricas de los beneficios y limitaciones de un método ágil específico.
- Identificar cualquier laguna en las investigaciones actuales a fin de sugerir áreas para una mayor investigación.
- Proporcionar un marco/antecedentes para situar adecuadamente las nuevas actividades de investigación.

Sin embargo, también se pueden realizar revisiones sistemáticas de la literatura para examinar en qué medida las pruebas empíricas apoyan/contradicen las hipótesis teóricas, o incluso para ayudar a generar nuevas hipótesis (véase, por ejemplo, [14]).

2.2 La importancia de las revisiones sistemáticas de la literatura

La mayoría de las investigaciones comienzan con una revisión de la literatura de algún tipo. Sin embargo, a menos que una revisión de la literatura sea completa y justa, tiene poco valor científico. Esta es la razón principal para llevar a cabo revisiones sistemáticas. Una revisión sistemática sintetiza el trabajo existente de una manera que es justa y se ve como tal. Por ejemplo, las revisiones sistemáticas deben realizarse de acuerdo con una estrategia de búsqueda predefinida. La estrategia de búsqueda debe permitir evaluar la integridad de la búsqueda. En particular, los investigadores que realicen un examen sistemático deben hacer todo lo posible por identificar y comunicar las investigaciones que no respalden su hipótesis de investigación preferida, así como identificar y comunicar las investigaciones que las respalden.

> *"De hecho, una de mis mayores quejas sobre el campo de la informática es que mientras que Newton podía decir, "Si he* ***visto un poco más lejos que otros, es porque me he parado sobre los hombros de gigantes", yo me veo obligado a*** *decir, "Hoy nos paramos sobre los pies* ***de cada uno****"."Tal vez el problema central que enfrentamos en toda la informática es cómo llegar a la situación en la que construimos sobre el trabajo de otros en lugar de rehacer tanto de una manera trivialmente diferente. Se supone que la ciencia es*

acumulativa, no una duplicación casi interminable del mismo tipo de cosas".

Richard Hamming 1968 Conferencia del premio Turning Award

Las revisiones sistemáticas de la literatura en todas las disciplinas nos permiten pararnos en los hombros de los gigantes y en la computación, nos permiten ponernos de pie unos a otros.

2.3 Ventajas y desventajas

Las ventajas de las revisiones sistemáticas de la literatura son que

- La metodología bien definida hace menos probable que los resultados de la literatura sean sesgados, aunque no protege contra el sesgo de publicación en los estudios primarios.
- Pueden proporcionar información sobre los efectos de algún fenómeno en una amplia gama de entornos y métodos empíricos. Si los estudios dan resultados coherentes, los exámenes sistemáticos proporcionan pruebas de que el fenómeno es robusto y transferible. Si los estudios dan resultados inconsistentes, pueden estudiarse las fuentes de variación.
- En el caso de los estudios cuantitativos, es posible combinar los datos mediante técnicas meta-analíticas. Esto aumenta la probabilidad de detectar efectos reales que los estudios individuales más pequeños no pueden detectar.

La principal desventaja de las revisiones sistemáticas de la literatura es que requieren un esfuerzo considerablemente mayor que las revisiones tradicionales de la literatura. Además, el aumento de la potencia de los meta-análisis también puede ser una desventaja, ya que es posible detectar pequeños sesgos así como efectos reales.

2.4 Características de las revisiones sistemáticas de la literatura

Algunas de las características que diferencian un examen sistemático de un examen convencional de la literatura de expertos son:

- Los exámenes sistemáticos comienzan por definir un protocolo de examen que especifica la cuestión de investigación que se está abordando y los métodos que se utilizarán para realizar el examen.
- Las revisiones sistemáticas se basan en una estrategia de búsqueda definida que tiene por objeto detectar la mayor cantidad posible de literatura pertinente.
- Las revisiones sistemáticas documentan su estrategia de búsqueda para que los lectores puedan evaluar su rigor y la exhaustividad y repetibilidad del proceso (teniendo en cuenta que las búsquedas en las bibliotecas digitales son casi imposibles de reproducir).
- Los exámenes sistemáticos requieren criterios explícitos de inclusión y exclusión para evaluar cada posible estudio primario.
- En los exámenes sistemáticos se especifica la información que se ha de obtener de cada estudio primario, incluidos los criterios de calidad con los que se ha de evaluar cada estudio primario.
- Un examen sistemático es un requisito previo para el meta-análisis cuantitativo.

2.5 Otros tipos de revisión

Hay otros dos tipos de revisión que complementan las revisiones sistemáticas de la literatura: los estudios cartográficos sistemáticos y las revisiones terciarias.

2.5.1 Estudios de mapeo sistemático

Si durante el examen inicial de un dominio antes de encargar un examen sistemático se descubre que es probable que existan muy pocas pruebas o que el tema sea

muy amplio, entonces un estudio de mapeo sistemático puede ser un ejercicio más apropiado que un examen sistemático.

Un estudio cartográfico sistemático permite trazar la evidencia en un dominio a un alto nivel de granularidad. Esto permite identificar grupos de pruebas y desiertos de pruebas para orientar el enfoque de futuros exámenes sistemáticos e identificar áreas para que se realicen más estudios primarios. En la sección 8 figura un esbozo del proceso de estudio de la cartografía sistemática en el que se destacan las principales diferencias con el proceso estándar de examen sistemático.

2.5.2 Revisiones Terciarias

En un ámbito en el que ya existen varios exámenes sistemáticos, tal vez sea posible realizar un examen terciario, que es un examen sistemático de los exámenes sistemáticos, a fin de responder a preguntas de investigación más amplias. Un examen terciario utiliza exactamente la misma metodología que un examen sistemático estándar de la literatura. Potencialmente requiere menos recursos que la realización de un nuevo examen sistemático de estudios primarios, pero depende de que se disponga de suficientes exámenes sistemáticos de alta calidad. El protocolo que se presenta en el Apéndice 3 es un protocolo para una revisión terciaria.

3. La ingeniería de software basada en la evidencia en el contexto

Es importante entender la relación de la Ingeniería de Software con otros dominios con respecto a la aplicabilidad del paradigma basado en la evidencia. Al hacerlo, podemos identificar cómo los procedimientos adoptados de otras disciplinas (particularmente la medicina) necesitan ser adaptados para adecuarse a la investigación y la práctica de la ingeniería de software.

Budgen y otros [6] entrevistaron a profesionales de diversos ámbitos que utilizan enfoques de investigación basados en pruebas y compararon sus prácticas de investigación con las de la ingeniería de programas informáticos. En el cuadro 1 se muestran los resultados de su evaluación de la similitud entre las prácticas de investigación de la ingeniería de programas informáticos y las de otros dominios. Muestra que la ingeniería de software es mucho más similar a las ciencias sociales que a la medicina. Esta similitud se debe a las prácticas experimentales, los tipos de materias y los procedimientos de cegamiento. Dentro de la Ingeniería de Software es difícil llevar a cabo ensayos controlados aleatorios o realizar un doble cegamiento. Además, la experiencia humana y el sujeto humano afectan al resultado de los experimentos.

Tabla 1 Comparación de la metodología experimental de la Ingeniería de Software con la de otras disciplinas

Disciplina	**Comparación con el SE (1 es un acuerdo perfecto, 0 es completo desacuerdo)**
Enfermería y partería	0.83
Atención Primaria	0.33
Química orgánica	0.83
Psicología empírica	0.66
Medicina clínica	0.17

Educación	0.83

Estos factores significan que la ingeniería de programas informáticos es significativamente diferente del ámbito médico tradicional en el que se desarrollaron por primera vez las revisiones sistemáticas. Por esto

razón por la cual hemos revisado estas directrices para incorporar ideas recientes del área de las ciencias sociales ([25], [11]). Además, la elección de las referencias en las que basar estas directrices se basó en nuestras discusiones con los investigadores de estas disciplinas.

4. El proceso de revisión

Una revisión sistemática de la literatura implica varias actividades discretas. Las directrices existentes para los exámenes sistemáticos tienen sugerencias ligeramente diferentes sobre el número y el orden de las actividades (véase el apéndice 1). Sin embargo, las directrices médicas y los libros de texto de sociología coinciden en general en las principales etapas del proceso.

Este documento resume las etapas de un examen sistemático en tres fases principales: Planificación del examen, realización del examen, informe del examen.

Las etapas asociadas a la *planificación del examen* son:
- Identificación de la necesidad de una revisión (Véase la sección 5.1).
- Encargar un examen (véase la sección 5.2).
- Especificando la(s) pregunta(s) de investigación (Véase la sección 5.3).
- Elaboración de un protocolo de revisión (véase la sección 5.4).
- Evaluar el protocolo de revisión (Ver sección 5.5).

Las etapas asociadas a la *realización del examen* son
- Identificación de las investigaciones (véase la sección 6.1).
- Selección de estudios primarios (Véase la sección 6.2).
- Evaluación de la calidad del estudio (véase la sección 6.3).
- Extracción de datos y vigilancia (véase la sección 6.4).
- Síntesis de datos (véase la sección 6.5).

Las etapas asociadas a la *presentación de informes sobre el examen* son
- Especificar los mecanismos de difusión (véase la sección 7.1).
- Formato del informe principal (véase la sección 7.2).
- Evaluación del informe (véase la sección 7.3).

Consideramos que todas las etapas anteriores son obligatorias excepto:
- Encargar un examen que depende de si el examen sistemático se realiza o no sobre una base comercial.
- La evaluación del protocolo de examen (5.5) y la evaluación del informe (7.3) son opcionales y dependen de los procedimientos de garantía de calidad decididos por el equipo de examen sistemático (y cualquier otra parte interesada).

Las etapas enumeradas anteriormente pueden parecer secuenciales, pero es importante reconocer que muchas de ellas implican una iteración. En particular, muchas actividades se inician durante la etapa de elaboración del protocolo, y se perfeccionan cuando tiene lugar la revisión propiamente dicha. Por ejemplo
- La selección de los estudios primarios se rige por criterios de inclusión y exclusión. Estos criterios se especifican inicialmente al redactar el protocolo, pero pueden perfeccionarse una vez definidos los criterios de calidad.
- Los formularios de extracción de datos preparados inicialmente durante la

construcción del protocolo se modificarán cuando se acuerden los criterios de calidad.

- Los métodos de síntesis de datos definidos en el protocolo podrán modificarse una vez que se hayan recogido los datos.

La hoja de ruta de revisiones sistemáticas preparada por el Grupo de Revisiones Sistemáticas de Berkeley demuestra muy claramente la naturaleza iterativa del proceso de revisión sistemática [24].

5. Planificación

Antes de emprender un examen sistemático es necesario confirmar la necesidad de dicho examen. En algunas circunstancias se encargan exámenes sistemáticos y en esos casos es necesario redactar un documento de encargo. Sin embargo, las actividades de examen previo más importantes son la definición de las preguntas de investigación que el examen sistemático abordará y la elaboración de un protocolo de examen (es decir, un plan) en el que se definan los procedimientos básicos de examen. El protocolo de examen también debe ser objeto de un proceso de evaluación independiente. Esto es particularmente importante para un examen encargado.

5.1 La necesidad de un examen sistemático

La necesidad de un examen sistemático surge de la exigencia de los investigadores de resumir toda la información existente sobre algún fenómeno de manera exhaustiva e imparcial. Esto puede ser para sacar conclusiones más generales sobre algún fenómeno que las que se desprenden de estudios individuales, o puede realizarse como preludio de nuevas actividades de investigación.

Ejemplos

Kitchenham y otros [21] sostuvieron que la estimación precisa de los costos es importante para la industria de los programas informáticos; que los modelos de estimación precisa de los costos se basan en datos de proyectos anteriores; que muchas empresas no pueden reunir suficientes datos para construir sus propios modelos. Por lo tanto, es importante saber si los modelos desarrollados a partir de los depósitos de datos pueden utilizarse para predecir los costos en una empresa concreta. Observaron que en varios estudios se ha abordado esa cuestión, pero que se han llegado a conclusiones diferentes. Llegaron a la conclusión de que es necesario determinar si, o en qué condiciones, los modelos derivados de los depósitos de datos pueden servir de apoyo a la estimación en una empresa concreta.

Jørgensen [17] señaló que, a pesar de que la mayoría de las investigaciones sobre estimación de costos de programas informáticos se concentran en modelos formales de estimación de costos y de que un gran número de administradores de tecnología de la información conocen las herramientas que aplican los modelos formales, la mayoría de las estimaciones de costos industriales se basan en el juicio de los expertos. Sostuvo que los investigadores necesitan saber si los profesionales de los programas informáticos son simplemente irracionales, o si el juicio de los expertos es tan exacto como los modelos formales o tiene otras ventajas que lo hacen más aceptable que los modelos formales.

En ambos casos los autores habían realizado investigaciones en el área temática y tenían conocimiento de primera mano de las cuestiones de investigación.

Antes de emprender un examen sistemático, los investigadores deben asegurarse de que es necesario un examen sistemático. En particular, los investigadores deben identificar y examinar todo examen sistemático existente del fenómeno de interés con arreglo a criterios de evaluación apropiados. El CRD [19] sugiere la siguiente lista de verificación:

- ¿Cuáles son los objetivos de la revisión?
- ¿Qué fuentes se buscaron para identificar los estudios primarios? ¿Hubo alguna restricción?

- ¿Cuáles fueron los criterios de inclusión/exclusión y cómo se aplicaron?
- ¿Qué criterios se utilizaron para evaluar la calidad de los estudios primarios?
- ¿Cómo se aplicaron los criterios de calidad?
- ¿Cómo se extrajeron los datos de los estudios primarios?
- ¿Cómo se sintetizaron los datos?
- ¿Cómo se investigaron las diferencias entre los estudios?
- ¿Cómo se combinaron los datos?
- ¿Fue razonable combinar los estudios?
- ¿Las conclusiones se derivan de las pruebas?

Los criterios de la Base de Datos de Resúmenes de Revisiones de Efectos (DARE) del CRD (http://www.york.ac.uk/inst/crd/crddatabases.htm#DARE) son aún más sencillos. Se basan en cuatro preguntas:

1. ¿Se describen y son adecuados los criterios de inclusión y exclusión del examen?
2. ¿Es probable que la búsqueda bibliográfica haya abarcado todos los estudios pertinentes?
3. ¿Los revisores evaluaron la calidad/validez de los estudios incluidos?
4. ¿Se describieron adecuadamente los datos/estudios básicos?

Ejemplos

Aplicamos el criterio DARE tanto al estudio de Kitchenham y otros [21] como al de Jørgensen [17]. Le dimos al estudio de Kitchenham y otros una puntuación de 4 y al de Jørgensen una puntuación de 3,5.
En el apéndice 2 se enumeran otros estudios puntuados utilizando los criterios DARE.

Desde un punto de vista más general, Greenlaugh [12] sugiere las siguientes preguntas:

- ¿Puede encontrar una pregunta clínica importante, que la revisión abordó? (Claramente, en la ingeniería de software, esto debería adaptarse para referirse a una pregunta importante de ingeniería de software).
- ¿Se realizó una búsqueda exhaustiva en las bases de datos apropiadas y se exploraron otras fuentes potencialmente importantes?
- ¿Se evaluó la calidad metodológica y se ponderaron los ensayos en consecuencia?
- ¿Qué tan sensibles son los resultados a la forma en que se ha hecho la revisión?
- ¿Se han interpretado los resultados numéricos con sentido común y teniendo en cuenta los aspectos más amplios del problema?

5.2 Encargar una revisión

A veces una organización requiere información sobre un tema específico pero no tiene el tiempo o la experiencia para realizar una literatura sistemática por sí misma. En tales casos, encargará a los investigadores que realicen una revisión sistemática de la literatura sobre el tema. Cuando esto ocurre, la organización debe producir un documento de encargo que especifique el trabajo requerido.

Un documento de encargo contendrá o considerará los siguientes puntos (adaptado de las directrices del CRD [12])

- Título del proyecto
- Antecedentes
- Preguntas de revisión

- Membresía del Grupo Asesor/Directivo (Investigadores, Profesionales, Miembros laicos, Creadores de políticas, etc.)
- Métodos del examen
- Calendario del proyecto
- Estrategia de difusión
- Infraestructura de apoyo
- Presupuesto
- Referencias

El documento de encargo puede utilizarse tanto para solicitar ofertas a los grupos de investigación que deseen realizar el examen como para servir de documento de orientación al grupo consultivo a fin de garantizar que el examen siga siendo centrado y pertinente en el contexto.

La fase de encargo de un examen sistemático no es necesaria para que un equipo de investigación realice un examen para sus propias necesidades o para que lo haga un estudiante de doctorado. Si no se lleva a cabo la fase de encargo, la estrategia de difusión debe incorporarse al protocolo de examen. Hasta el momento no hay ejemplos de SLR encargados en el ámbito de la ingeniería de programas informáticos.

5.3 La(s) pregunta(s) de investigación

La especificación de las preguntas de investigación es la parte más importante de cualquier revisión sistemática. Las preguntas de revisión dirigen toda la metodología de revisión sistemática:

- El proceso de búsqueda debe identificar los estudios primarios que abordan las cuestiones de investigación.
- El proceso de extracción de datos debe extraer los elementos de datos necesarios para responder a las preguntas.
- El proceso de análisis de datos debe sintetizar los datos de tal manera que las preguntas puedan ser respondidas.

5.3.1 Tipos de preguntas

La actividad más importante durante la planificación es formular la(s) pregunta(s) de investigación. En las directrices australianas sobre RMN [1] se identifican seis tipos de preguntas sobre la atención de la salud que pueden abordarse mediante exámenes sistemáticos:

1. Evaluar el efecto de la intervención.
2. Evaluar la frecuencia o el ritmo de una condición o enfermedad.
3. Determinar el rendimiento de una prueba de diagnóstico.
4. Identificar la etiología y los factores de riesgo.
5. Identificar si una condición puede ser predicha.
6. Evaluar el valor económico de una intervención o procedimiento.

En la ingeniería de programas informáticos, no está claro cuál sería el equivalente de una prueba de diagnóstico, pero las demás preguntas pueden adaptarse a las cuestiones de ingeniería de programas informáticos de la siguiente manera:

- Evaluar el efecto de una tecnología de ingeniería de software.
- Evaluar la frecuencia o el ritmo de un factor de desarrollo de un proyecto, como la adopción de una tecnología, o la frecuencia o el ritmo de éxito o

fracaso de un proyecto.

- Identificar el costo y los factores de riesgo asociados a una tecnología.
- Identificar el impacto de las tecnologías en los modelos de fiabilidad, rendimiento y coste.

- Análisis de costo-beneficio del empleo de tecnologías específicas de desarrollo de software o aplicaciones de software.

Las directrices médicas a menudo proporcionan diferentes pautas y procedimientos para diferentes tipos de preguntas. Este documento no llega a este nivel de detalle.

La cuestión crítica en cualquier revisión sistemática es hacer la pregunta correcta. En este contexto, la pregunta correcta suele ser aquella que:

- Es significativo e importante tanto para los profesionales como para los investigadores. Por ejemplo, a los investigadores podría interesarles saber si una técnica de análisis específica conduce a una estimación significativamente más precisa de los defectos restantes después de las inspecciones del diseño. Sin embargo, un profesional podría querer saber si la adopción de una técnica de análisis específica para predecir los defectos restantes es más eficaz que la opinión de los expertos en la identificación de los documentos de diseño que requieren una nueva inspección.
- conducirá a cambios en la práctica actual de la ingeniería de software o a una mayor confianza en el valor de la práctica actual. Por ejemplo, a los investigadores y profesionales les gustaría saber en qué condiciones un proyecto puede adoptar con seguridad tecnologías ágiles y en qué condiciones no debería hacerlo.
- Identificará las discrepancias entre las creencias comunes y la realidad.

No obstante, hay exámenes sistemáticos que plantean preguntas que interesan principalmente a los investigadores. En esos exámenes se formulan preguntas que identifican y/o amplían las futuras actividades de investigación. Por ejemplo, un examen sistemático de una tesis de doctorado debería identificar la base existente para el trabajo del estudiante de investigación y dejar claro dónde encaja la investigación propuesta en el cuerpo de conocimientos actual.

Ejemplos

Kitchenham y otros [21] tenían tres preguntas de investigación:

Pregunta 1: ¿Qué pruebas hay de que los modelos de estimación entre empresas no difieren significativamente de los modelos de estimación dentro de la empresa para predecir el esfuerzo de los proyectos de software/Web?

Pregunta 2: ¿Qué características de los conjuntos de datos del estudio y los métodos de análisis de datos utilizados en el estudio afectan al resultado de los estudios de precisión de la estimación de los esfuerzos internos y externos?

Pregunta 3: ¿Qué procedimiento experimental es el más apropiado para los estudios que comparan modelos de estimación dentro de una misma empresa y entre empresas?

Jørgensen [17] tenía dos preguntas de investigación:

1. ¿Deberíamos esperar estimaciones de esfuerzo más precisas al aplicar el juicio de los expertos o los modelos?
2. ¿Cuándo deberían basarse las estimaciones de los esfuerzos de desarrollo de programas informáticos en el juicio de los expertos, cuándo en modelos y cuándo en una combinación de juicio de expertos y modelos?

En ambos casos, los autores eran conscientes, por investigaciones anteriores, de que los resultados eran mixtos, por lo que en cada caso añadieron una pregunta destinada a investigar las condiciones en las que se obtienen resultados diferentes.

5.3.2 Estructura de la pregunta

Las directrices médicas recomiendan considerar una pregunta sobre la eficacia de un tratamiento desde tres puntos de vista:

- La población, es decir, las personas afectadas por la intervención.

- Las intervenciones, que suelen ser una comparación entre dos o más tratamientos alternativos.
- Los resultados, es decir, los factores clínicos y económicos que se utilizarán para comparar las intervenciones.

Más recientemente, Petticrew y Roberts sugieren utilizar los criterios PICOC (Población, Intervención, Comparación, Resultado, Contexto) para enmarcar las preguntas de investigación [25]. Estos criterios amplían las directrices médicas originales con:
Comparación: Es decir, qué es la intervención que se compara con el contexto: es decir, cuál es el contexto en el que se realiza la intervención.

Además, se pueden identificar los diseños de estudio apropiados para responder a las preguntas de examen y utilizarlos para orientar la selección de los estudios primarios.

Discutimos estos criterios desde el punto de vista de la ingeniería de software a continuación.

Población
En los experimentos de ingeniería de software, las poblaciones podrían ser cualquiera de las siguientes:
- Un papel específico de ingeniería de software, por ejemplo, probadores, gerentes.
- Una categoría de ingeniero de software, por ejemplo, un ingeniero novato o experimentado.
- Un área de aplicación, por ejemplo, sistemas informáticos, sistemas de mando y control.
- Un grupo industrial como las empresas de telecomunicaciones o las pequeñas empresas de informática.

Una pregunta puede referirse a grupos de población muy específicos, por ejemplo, probadores novatos o arquitectos de software con experiencia que trabajan en sistemas informáticos. En medicina, las poblaciones se definen para reducir el número de estudios primarios prospectivos. En la ingeniería de programas informáticos se realizan muchos menos estudios primarios, por lo que tal vez sea necesario evitar toda restricción de la población hasta que se llegue a considerar las consecuencias prácticas del examen sistemático.

Intervención
La intervención es la metodología/herramienta/tecnología/procedimiento de software que aborda una cuestión específica, por ejemplo, las tecnologías para realizar tareas concretas como la especificación de requisitos, las pruebas de sistemas o la estimación de los costos de software.

Comparación
Esta es la metodología/herramienta/tecnología/procedimiento de ingeniería de software con la que se está comparando la intervención. Cuando la tecnología de comparación es la tecnología convencional o de uso común, se suele denominar tratamiento de "control". La situación de control debe describirse adecuadamente. En particular, "no utilizar la intervención" es inadecuado como descripción del

tratamiento de control. Las técnicas de ingeniería de programas informáticos suelen requerir capacitación. Si se compara a las personas que utilizan una técnica con las que no la utilizan, el efecto de la técnica se confunde con el efecto de la capacitación. Es decir, cualquier efecto puede deberse a la capacitación y no a la técnica específica. Este es un problema particular si los participantes son estudiantes.

Resultados

Los resultados deben estar relacionados con factores de importancia para los profesionales, como una mayor fiabilidad, la reducción de los costos de producción y la reducción del tiempo de comercialización. Todos los resultados pertinentes

se deben especificar los resultados. Por ejemplo, en algunos casos se requieren intervenciones que mejoren algún aspecto de la producción de programas informáticos sin afectar a otro, por ejemplo, una mayor fiabilidad sin aumentar el costo.

Un problema particular de los experimentos de ingeniería de programas informáticos es el uso generalizado de medidas sustitutivas, por ejemplo, los defectos encontrados durante las pruebas del sistema como sustituto de la calidad, o las medidas de acoplamiento para la calidad del diseño. Los estudios que utilizan medidas sustitutivas pueden ser engañosos y las conclusiones basadas en esos estudios pueden ser menos sólidas.

Contexto
En el caso de la ingeniería de programas informáticos, este es el contexto en el que se realiza la comparación (por ejemplo, el mundo académico o la industria), los participantes que intervienen en el estudio (por ejemplo, profesionales, académicos, consultores, estudiantes) y las tareas que se realizan (por ejemplo, a pequeña o gran escala). Muchos experimentos con programas informáticos tienen lugar en el mundo académico utilizando participantes estudiantes y tareas en pequeña escala. Es poco probable que esos experimentos sean representativos de lo que podría ocurrir con los profesionales que trabajan en la industria. En algunos exámenes sistemáticos se podría optar por excluir esos experimentos, aunque en el ámbito de la ingeniería de programas informáticos, tal vez sean el único tipo de estudios disponibles.

Diseños experimentales
En los estudios médicos, los investigadores pueden restringir las revisiones sistemáticas a los estudios primarios de un tipo particular. Por ejemplo, las revisiones Cochrane suelen restringirse a ensayos controlados aleatorios (ECA). En otras circunstancias, la naturaleza de la cuestión y el tema central que se aborda puede sugerir que ciertos diseños de estudio son más apropiados que otros. Sin embargo, este enfoque sólo puede adoptarse en una disciplina en la que el gran número de trabajos de investigación es un problema importante. En la ingeniería de programas informáticos, es más probable que la escasez de estudios primarios sea el problema de los exámenes sistemáticos y que necesitemos protocolos para agregar información de estudios de tipos muy diferentes.

Ejemplos

Kitchenham y otros[21] utilizaron los criterios de la PICO y definieron los elementos de la pregunta como
Población: software o proyecto web.
Intervención: modelo de estimación de esfuerzo de proyectos interempresariales. **Comparación: modelo de estimación del esfuerzo de** un proyecto de una sola empresa **Resultados: precisión de la predicción** o la estimación**.**

Jørgensen [17] no usó una versión estructurada de sus preguntas de investigación.

5.4 Elaboración de un protocolo de examen

En un protocolo de examen se especifican los métodos que se utilizarán para llevar a cabo un examen sistemático específico. Es necesario un protocolo

predefinido para reducir la posibilidad de sesgo del investigador. Por ejemplo, sin un protocolo, es posible que la selección de los estudios individuales o el análisis se vean impulsados por las expectativas del investigador. En la medicina, los protocolos de examen suelen someterse a un examen por homólogos.

Los componentes de un protocolo incluyen todos los elementos del examen más alguna información adicional de planificación:

- Antecedentes. El fundamento de la encuesta.
- Las preguntas de investigación que la revisión pretende responder.
- La estrategia que se utilizará para buscar estudios primarios, incluidos los términos de búsqueda y los recursos que se buscarán. Los recursos incluyen bibliotecas digitales, revistas específicas y actas de conferencias. Un estudio inicial de mapeo puede ayudar a determinar una estrategia apropiada.
- Criterios de selección del estudio. Los criterios de selección de estudios se utilizan para determinar qué estudios se incluyen o excluyen de un examen sistemático. Suele ser útil pilotar los criterios de selección en un subconjunto de estudios primarios.
- Procedimientos de selección de estudios. El protocolo debe describir cómo se aplicarán los criterios de selección, por ejemplo, cuántos asesores evaluarán cada estudio primario prospectivo y cómo se resolverán los desacuerdos entre los asesores.
- Listas de control y procedimientos de evaluación de la calidad de los estudios. Los investigadores deben desarrollar listas de control de calidad para evaluar los estudios individuales. El propósito de la evaluación de calidad guiará el desarrollo de las listas de control.
- Estrategia de extracción de datos. Esto define cómo se obtendrá la información requerida de cada estudio primario. Si los datos requieren manipulación o que se hagan suposiciones e inferencias, el protocolo debe especificar un proceso de validación apropiado.
- Síntesis de los datos extraídos. Esto define la estrategia de síntesis. Esto debería aclarar si se pretende realizar un meta-análisis formal y, en caso afirmativo, qué técnicas se utilizarán.
- Estrategia de difusión (si no está ya incluida en un documento de encargo).
- Calendario del proyecto. Esto debería definir el calendario de revisión.

En el apéndice 3 figura un ejemplo de protocolo para un examen terciario. Se trata de un estudio sencillo, por lo que el protocolo es bastante corto. En nuestra experiencia, los protocolos pueden ser documentos muy largos. En este caso, el protocolo es corto porque el proceso de búsqueda es relativamente limitado y los procesos de extracción y análisis de datos son relativamente sencillos.

5.5 Evaluación de un protocolo de revisión

El protocolo es un elemento crítico de cualquier revisión sistemática. Los investigadores deben acordar un procedimiento para evaluar el protocolo. Si se dispone de fondos adecuados, se debe pedir a un grupo de expertos independientes que revise el protocolo. Posteriormente se puede pedir a los mismos expertos que revisen el informe final.

Los estudiantes de doctorado deben presentar su protocolo a sus supervisores para que lo revisen y lo critiquen.

Las preguntas básicas de revisión del SLR que se tratan en la sección 5.1 pueden adaptarse para ayudar a la evaluación de un protocolo de revisión sistemática. Además, se puede comprobar la coherencia interna del protocolo para confirmarlo:

- Las cadenas de búsqueda se derivan adecuadamente de las preguntas de investigación.

- Los datos que se extraigan abordarán adecuadamente la(s) cuestión(es) de la investigación.
- El procedimiento de análisis de datos es apropiado para responder a las preguntas de la investigación.

5.6 Lecciones aprendidas para la construcción de protocolos

Brereton y otros [5] identifican una serie de cuestiones que los investigadores deben anticipar durante la construcción del protocolo:

- Un estudio de mapeo previo puede ayudar a determinar el alcance de las preguntas de investigación.
- Espere revisar las preguntas durante la elaboración del protocolo, a medida que aumente la comprensión del problema.
- Todos los miembros del equipo de examen sistemático deben participar activamente en la elaboración del protocolo de examen, para que comprendan cómo realizar el proceso de extracción de datos.
- Pilotaje del protocolo de investigación es esencial. Encontrará errores en los procedimientos de recopilación y agregación de datos. También puede indicar la necesidad de modificar la metodología destinada a abordar las cuestiones de investigación, incluida la modificación de los formularios de extracción de datos y los métodos de síntesis.

Staples y Niazi [27] recomiendan limitar el alcance de una literatura sistemática eligiendo preguntas de investigación claras y estrechas.

6. La realización de la revisión

Una vez que el protocolo haya sido acordado, la revisión propiamente dicha puede comenzar. Sin embargo, como se ha señalado anteriormente, los investigadores deben esperar probar cada uno de los pasos descritos en esta sección cuando construyan su protocolo de investigación.

6.1 Identificación de la investigación

El objetivo de un examen sistemático es encontrar el mayor número posible de estudios primarios relacionados con la cuestión de la investigación mediante una estrategia de búsqueda imparcial. El rigor del proceso de búsqueda es uno de los factores que distingue las revisiones sistemáticas de las tradicionales.

6.1.1 Generando una estrategia de búsqueda

Es necesario determinar y seguir una estrategia de búsqueda. Ésta debe elaborarse en consulta con los bibliotecarios u otras personas con experiencia pertinente. Las estrategias de búsqueda suelen ser iterativas y se benefician de ellas:

- Las búsquedas preliminares tienen por objeto tanto identificar los exámenes sistemáticos existentes como evaluar el volumen de los estudios potencialmente pertinentes.
- Búsquedas de ensayos utilizando varias combinaciones de términos de búsqueda derivados de la pregunta de investigación.
- Comprobar las cadenas de investigación de los ensayos con las listas de estudios primarios ya conocidos.
- Consultas con expertos en la materia.

Un enfoque general consiste en desglosar la cuestión en facetas individuales, es decir, población, intervención, comparación, resultados, contexto y diseños de estudio, como se expone en la sección 5.3.2. A continuación, elaborar una lista de sinónimos, abreviaturas y ortografías alternativas. Otros términos pueden obtenerse considerando los encabezamientos temáticos utilizados en las revistas y bases de

datos. Se pueden construir sofisticadas cadenas de búsqueda utilizando ANDs y ORs booleanos.

Las búsquedas iniciales de estudios primarios pueden realizarse utilizando bibliotecas digitales, pero esto no es suficiente para una revisión sistemática completa. También deben buscarse otras fuentes de pruebas (a veces manualmente), entre ellas:

- Listas de referencia de estudios primarios y artículos de revisión pertinentes
- Revistas (incluidas las revistas de las empresas, como el IBM Journal of Research and Development), literatura gris (es decir, informes técnicos, trabajos en curso) y actas de conferencias
- Registros de investigación
- La Internet.

También es importante identificar investigadores específicos a los que dirigirse directamente para obtener asesoramiento sobre el material de origen apropiado.

Los investigadores médicos han desarrollado estrategias de búsqueda preempaquetadas. Los investigadores en ingeniería de programas informáticos deben elaborar y publicar esas estrategias, incluida la identificación de las bibliotecas digitales pertinentes.

Un problema para los SLR de ingeniería de software es que puede haber relativamente pocos estudios sobre un tema en particular. En esos casos puede ser una buena idea buscar estudios en disciplinas conexas, por ejemplo, la sociología para las prácticas de trabajo en grupo, y la psicología para el diseño de notación y/o los enfoques de solución de problemas.

Ejemplo

Jørgensen [16] investigó cuándo se puede esperar que las estimaciones de los expertos tengan una exactitud aceptable en comparación con los modelos formales, revisando los estudios pertinentes sobre el juicio humano (por ejemplo, los estudios de estimación del tiempo) y comparando sus resultados con los de los estudios de ingeniería de programas informáticos.

6.1.2 Sesgo de publicación

El sesgo de publicación se refiere al problema de que es más probable que se publiquen resultados *positivos* que *negativos*. El concepto de resultados *positivos* o *negativos depende a* veces del punto de vista del investigador. (Por ejemplo, la evidencia de que las mastectomías completas no siempre se requerían para el cáncer de mama era en realidad un resultado extremadamente positivo para los enfermos de cáncer de mama).

Sin embargo, el sesgo de publicación sigue siendo un problema, en particular para los experimentos formales, en los que el hecho de no rechazar la hipótesis nula se considera menos interesante que un experimento capaz de rechazar la hipótesis nula. El sesgo de publicación es aún más problemático cuando los métodos/técnicas están patrocinados por grupos influyentes de la industria del software. Por ejemplo, el Ministerio de Defensa de los Estados Unidos es una organización extremadamente importante e influyente que patrocinó el desarrollo del Modelo de Madurez de la Capacidad y utilizó su influencia para alentar a la industria a adoptar el MMC. En tales circunstancias, pocas empresas querrían publicar resultados negativos y existe un fuerte incentivo para publicar artículos que apoyen el nuevo método/técnica.

El sesgo de publicación puede dar lugar a un sesgo sistemático en los exámenes sistemáticos, a menos que se hagan esfuerzos especiales para abordar este problema. Muchas de las estrategias de búsqueda estándar identificadas

anteriormente se utilizan para abordar este problema, entre ellas:

- Escaneando la literatura gris
- Escaneando las actas de la conferencia

- Contactar con expertos e investigadores que trabajan en el área y preguntarles si conocen algún resultado no publicado.

Además, se pueden utilizar técnicas de análisis estadístico para determinar la posible importancia del sesgo de publicación (véase la sección 6.5.7).

6.1.3 Gestión de la bibliografía y recuperación de documentos

Los paquetes bibliográficos como Reference Manager o Endnote pueden ser útiles para gestionar el gran número de referencias que se pueden obtener de una búsqueda bibliográfica exhaustiva.

Una vez que se hayan finalizado las listas de referencia, será necesario obtener los artículos completos de los estudios potencialmente útiles. Se necesita un sistema de registro para asegurarse de que se obtengan todos los estudios pertinentes.

6.1.4 Documentar la búsqueda

El proceso de realizar un examen sistemático de la literatura debe ser transparente y reproducible (en la medida de lo posible):

- La revisión debe documentarse con suficiente detalle para que los lectores puedan evaluar la minuciosidad de la búsqueda.
- La búsqueda debe documentarse en el momento en que se produce y los cambios deben ser anotados y justificados.
- Los resultados de la búsqueda sin filtrar deben guardarse y conservarse para un posible nuevo análisis.

Los procedimientos para documentar el proceso de búsqueda se indican en el cuadro 2.

Cuadro 2 Documentación del proceso de búsqueda

Fuente de datos	Documentación
Biblioteca digital	Nombre de la base de datos Estrategia de búsqueda de la base de datos Fecha de la búsqueda Años cubiertos por la búsqueda
Búsquedas manuales en el diario	Nombre de la revista Años de búsqueda Cualquier problema que no se haya buscado
Las actas de la conferencia	Título del procedimiento Nombre de la conferencia (si es diferente) Traducción del título (si es necesario) Nombre de la revista (si se publica como parte de una revista)
Esfuerzos para identificar estudios no publicados	Grupos de investigación e investigadores contactados (Nombres y datos de contacto) Sitios web de investigación buscados (Fecha y URL)
Otras fuentes	Fecha de la búsqueda / URL contactada Cualquier condición específica relativa a la búsqueda

Los investigadores deben especificar su justificación para:

- Las bibliotecas digitales a buscar.
- Se buscará en la revista y en las actas de la conferencia.
- El uso de búsquedas electrónicas o manuales o una combinación de ambas.

Aunque la mayoría de los libros de texto hacen hincapié en el uso de procedimientos de búsqueda electrónica, no suelen ser suficientes por sí solos, y algunos investigadores abogan firmemente por el uso de búsquedas manuales (por ejemplo, Jørgensen, [18]).

6.1.5 Lecciones aprendidas para los procedimientos de búsqueda

Brereton y otros [5] identifican varias cuestiones que deben abordarse al especificar los procedimientos de búsqueda electrónica:

- Existen estrategias de búsqueda alternativas que permiten alcanzar diferentes tipos de criterios de finalización de la búsqueda. Debe seleccionar y justificar una estrategia de búsqueda que sea apropiada para su pregunta de investigación. Por ejemplo, saber la fecha de publicación del primer artículo sobre un tema específico restringe los años en los que se debe buscar. Además, si va a restringir su búsqueda a revistas y actas de conferencias específicas, debe justificarlo.
- Tenemos que buscar en muchas fuentes electrónicas diferentes; ninguna fuente única encuentra todos los estudios primarios.
- Los actuales motores de búsqueda de ingeniería de software no están diseñados para apoyar revisiones sistemáticas de la literatura. A diferencia de los investigadores médicos, los investigadores en ingeniería de software necesitan realizar búsquedas que dependen de los recursos.

En un intento de realizar una búsqueda exhaustiva, Brereton y otros [5] identificaron siete fuentes electrónicas de relevancia para los ingenieros de software:

- IEEExplore
- Biblioteca digital de ACM:
- Becario de Google (scholar.google.com)
- Biblioteca Citeseer (citeseer.ist.psu.edu)
- Inspec (www.iee.org/Publish/INSPEC/)
- ScienceDirect (www.sciencedirect.com)
- EI Compendex (www.engineeringvillage2.org/Controller/Servlet/AthensService).

Sin embargo, también puede ser necesario considerar la posibilidad de utilizar SpringerLink para acceder a revistas como Empirical Software Engineering y Springer Conference Proceedings, o SCOPUS (que afirma ser la mayor base de datos de resúmenes y citas).

Ejemplos

Kitchenham y otros [21] utilizaron sus preguntas estructuradas para construir cadenas de búsqueda para su uso con bases de datos electrónicas. Los sinónimos y ortografías alternativas identificadas para cada uno de los elementos de la pregunta y los vincularon usando el O booleano, por ejemplo:

Población: software O aplicación O producto O Web O WWW O Internet O World- Wide Web O proyecto O desarrollo

Intervención: cruzada de empresas O cruzada de organizaciones O cruzada de organizaciones O modelo de organización múltiple O modelo de organización múltiple O modelación O esfuerzo de modelación O costo O estimación de recursos O predicción O evaluación

Contraste: dentro de la organización o dentro de la organización o dentro de la organización o dentro de la organización o una sola empresa o una sola organización

Resultado: Exactitud O Magnitud Media Error Relativo

Las cadenas de búsqueda se construyeron enlazando las cuatro listas de OR usando el AND booleano.

Las cadenas de búsqueda se utilizaron en 6 bibliotecas digitales:

- INSPEC
- El Compendex
- Science Direct
- Web de la Ciencia

- IEEExplore
- Biblioteca digital de ACM

Era necesario adaptar las cadenas de búsqueda para que se ajustaran a los requisitos específicos de las diferentes bases de datos. Además, los investigadores buscaron en varias revistas individuales (J) y en fuentes de actas de conferencias (C):

- Ingeniería de Software Empírica (J)
- Tecnología de la información y el software (J)
- Mejora y práctica de los procesos de software (J)
- Ciencia de la gestión (J)
- Simposio Internacional sobre Métricas de Programas Informáticos (C)
- Conferencia Internacional sobre Ingeniería de Programas Informáticos (C)
- Evaluation and Assessment in Software Engineering (búsqueda manual)

(C) Se eligieron estas fuentes porque habían publicado documentos sobre el tema.

Además, Kitchenham y otros comprobaron las referencias de cada artículo pertinente y se dirigieron a los investigadores que publicaron sobre el tema para preguntarles si habían publicado (o estaban en proceso de publicar) algún otro artículo sobre el tema.

Jørgensen [17] utilizó una base de datos existente de artículos de revistas que había identificado para otra revisión (Jørgensen y Shepperd [15]). Jørgensen y Shepperd buscaron manualmente en todos los volúmenes de más de 100 revistas los artículos sobre la estimación del costo del software. Las revistas se identificaron leyendo las listas de referencia de los documentos de estimación de costos, buscando en Internet y la propia experiencia de los investigadores. Los artículos individuales se clasificaron y registraron en una base de datos de acceso público (www.simula.no\BESTweb.

Para los documentos de conferencias, Jørgensen buscó los documentos identificados por la base de datos INSPEC utilizando la siguiente cadena de búsqueda:

(estimación del esfuerzo o estimación del coste) y el desarrollo de software.

También se puso en contacto con los autores de los documentos pertinentes y se le informó de otro documento pertinente.

Kitchenham y otros utilizaron el procedimiento recomendado por la mayoría de las directrices para realizar una revisión sistemática. Sin embargo, resultó en cadenas de búsqueda extremadamente largas que necesitaban ser adaptadas a motores de búsqueda específicos. Jørgensen [17] utilizó una base de datos previamente construida para un amplio estudio de estimación de costos de software. Este es un ejemplo de lo valioso que puede ser un estudio cartográfico. También utilizó una cadena de búsqueda bastante simple en la base de datos INSPEC.
Kitchenham y otros intentaron producir una cadena de búsqueda muy específica para su pregunta de investigación, pero aún así encontraron un gran número de falsos positivos. En la práctica, una cadena de búsqueda más simple podría haber sido igual de efectiva.

Es importante señalar que ninguno de los dos estudios basó su proceso de búsqueda únicamente en la búsqueda en bibliotecas digitales. Ambos estudios tenían preguntas de investigación muy específicas y los investigadores eran conscientes de que el número de documentos que abordaban el tema sería pequeño. Por lo tanto, ambos estudios se esforzaron por realizar una búsqueda exhaustiva.

6.2 Selección del estudio

Una vez que se hayan obtenido los estudios primarios potencialmente pertinentes, es preciso evaluar su pertinencia real.

6.2.1 Criterios de selección del estudio

Los criterios de selección de los estudios tienen por objeto identificar los estudios primarios que aportan pruebas directas sobre la cuestión de la investigación. A fin de reducir la probabilidad de sesgo, los criterios de selección deben decidirse durante la definición del protocolo, aunque pueden perfeccionarse durante el proceso de búsqueda.

Los criterios de inclusión y exclusión deben basarse en la cuestión de la investigación. Deberían ser puestos a prueba para asegurar que puedan ser interpretados de manera fiable y que clasifiquen los estudios correctamente.

Ejemplos

Kitchenham y otros utilizaron los siguientes criterios de inclusión:

- cualquier estudio que comparara las predicciones de los modelos entre empresas con los modelos dentro de la empresa basados en el análisis de los datos de un proyecto de una sola empresa.

Utilizaron los siguientes criterios de exclusión:

- estudios en los que los proyectos se recogieron sólo de un pequeño número de fuentes diferentes (por ejemplo, 2 ó 3 empresas),
- estudios en los que se compararon modelos derivados de un conjunto de datos internos con predicciones de un modelo general de estimación de costos.

Jørgensen [17] incluyó documentos que comparan la estimación del esfuerzo de desarrollo de software basado en juicios y en modelos. También excluyó un documento pertinente debido a "información incompleta sobre cómo se derivaron las estimaciones".

Problemas:

- Las normas médicas señalan que es importante evitar, en la medida de lo posible, las exclusiones basadas en el lenguaje del estudio primario. Esto puede no ser tan importante para la Ingeniería de Software.
- Es posible que las decisiones de inclusión se vean afectadas por el conocimiento de los autores, las instituciones, las revistas o el año de publicación. Algunos investigadores médicos han sugerido que las revisiones deben hacerse después de que se haya eliminado esa información. Sin embargo, lleva tiempo hacerlo y las pruebas experimentales sugieren que enmascarar el origen de los estudios primarios no mejora las revisiones [4].

6.2.2 Proceso de selección del estudio

La selección del estudio es un proceso de varias etapas. Inicialmente, los criterios de selección deben interpretarse de manera liberal, de modo que, a menos que un estudio identificado por las búsquedas electrónicas y manuales pueda excluirse claramente sobre la base del título y el resumen, debe obtenerse una copia completa. Sin embargo, Brereton y otros [5] señalan que *"el nivel* de los resúmenes de ingeniería informática y de programas informáticos es demasiado bajo para poder confiar en él al seleccionar los estudios primarios. También debería revisar las conclusiones".

El siguiente paso es aplicar criterios de inclusión/exclusión basados en cuestiones prácticas [11] como:

- Idioma
- Revista
- Autores
- Configuración
- Participantes o sujetos
- Diseño de la investigación
- Método de muestreo
- Fecha de publicación.

Staples y Niazi señalan que a veces es necesario considerar las cuestiones que no se están abordando para perfeccionar sus criterios de exclusión [27].

Ejemplo

La pregunta de la investigación de Staples y Niazi fue

- ¿Por qué las organizaciones se embarcan en iniciativas de SPI basadas en la MMC?

También definieron las cuestiones de investigación complementarias que no se estaban investigando:

- ¿Qué motiva a los individuos a apoyar la adopción del SPI basado en la MMC en una organización?
- ¿Por qué las organizaciones deberían embarcarse en iniciativas de SPI basadas en la MMC?
- ¿Qué razones para embarcarse en un SPI basado en la MMC son las más importantes para las organizaciones?
- ¿Qué beneficios han recibido las organizaciones de las iniciativas de SPI basadas en la MMC?
- ¿Cómo deciden las organizaciones embarcarse en iniciativas de SPI basadas en la MMC?
- ¿Qué problemas tienen las organizaciones en el momento en que deciden adoptar el SPI basado en la MMC?

Esto clarificó los límites de su pregunta de investigación de interés, por ejemplo, se preocuparon por las motivaciones de las organizaciones no las motivaciones de los individuos y se preocuparon por el porqué las organizaciones rechazaron la MMC no por qué la adoptaron. Encontraron que este proceso mejoró directamente y aclaró su selección de estudios primarios y el proceso de extracción de datos.

A veces, los investigadores emprenden una tercera etapa del proceso de selección basada en criterios de calidad detallados.

La mayoría de los libros de texto de la SLR recomiendan mantener una lista de estudios excluidos que identifique la razón de la exclusión. Sin embargo, en nuestra experiencia, las búsquedas electrónicas iniciales dan como resultado un gran número de trabajos totalmente irrelevantes, es decir, trabajos que no sólo no abordan ningún aspecto de las cuestiones de investigación sino que ni siquiera tienen nada que ver con la ingeniería de software. Por lo tanto, recomendamos mantener una lista de documentos excluidos, sólo después de que los documentos totalmente irrelevantes hayan sido excluidos, en particular, manteniendo un registro de los estudios primarios candidatos que se excluyen como resultado de los criterios de inclusión/exclusión más detallados.

6.2.3 Fiabilidad de las decisiones de inclusión

Cuando dos o más investigadores evalúan cada trabajo, el acuerdo entre los investigadores puede ser medido usando la estadística de Cohen Kappa [9]. El valor inicial de las estadísticas Kappa debe ser documentado en el informe final. Cada desacuerdo debe ser discutido y resuelto. Puede tratarse de una cuestión de remitirse al protocolo o puede implicar escribir a los autores para obtener información adicional. La incertidumbre acerca de la inclusión/exclusión de algunos estudios debe investigarse mediante un análisis de sensibilidad.

Un solo investigador (como un estudiante de doctorado) debería considerar la posibilidad de examinar los documentos incluidos y excluidos con su asesor, un grupo de expertos u otros investigadores. Como alternativa, los investigadores individuales pueden aplicar un enfoque de prueba y reevaluar una muestra aleatoria de los estudios primarios encontrados después de la selección inicial para comprobar la coherencia de sus decisiones de inclusión/exclusión.

6.3 Evaluación de la calidad del estudio

Además de los criterios generales de inclusión/exclusión, se considera fundamental evaluar la "calidad" de los estudios primarios:

- Proporcionar criterios de inclusión/exclusión aún más detallados.
- Investigar si las diferencias de calidad explican las diferencias en los resultados de los estudios.

- Como un medio de ponderar la importancia de los estudios individuales cuando se están sintetizando los resultados.
- Para orientar la interpretación de los hallazgos y determinar la fuerza de las inferencias.
- Para orientar las recomendaciones para futuras investigaciones.

Una dificultad inicial es que no hay una definición acordada de la "calidad" del estudio. Sin embargo, tanto las Guías del CRD [19] como el Manual del Revisor Cochrane [7] sugieren que la calidad se relaciona con la medida en que el estudio minimiza el sesgo y maximiza la validez interna y externa (ver Tabla 3).

Cuadro 3 Definiciones de conceptos de calidad

Término	Sinónimos	Definición
Bias	Error sistemático	Una tendencia a producir resultados que se apartan sistemáticamente de los "verdaderos" resultados. Los resultados imparciales son válidos internamente
Validez interna	Validez	La medida en que el diseño y la realización del estudio pueden evitar el error sistemático. La validez interna es una requisito previo para la validez externa.
Validez externa	Generalización, aplicabilidad	La medida en que los efectos observados en el estudio son aplicables fuera del mismo.

La mayoría de las listas de control de calidad (véase la sección 6.3.2) incluyen preguntas destinadas a evaluar en qué medida los artículos han abordado el sesgo y la validez.

6.3.1 La jerarquía de la evidencia

Las directrices médicas sugieren que una evaluación inicial de la calidad puede basarse en el tipo de diseño de experimento que se utilice. Por lo tanto, podríamos calificar un ensayo controlado aleatorio como más fiable que un estudio de observación. Esto ha llevado al concepto de una jerarquía de pruebas con pruebas de revisiones sistemáticas y experimentos controlados aleatorios en la parte superior de la jerarquía y pruebas de cuasi-experimentos y opinión de expertos en la parte inferior de la jerarquía (véase [19] y [2]).
Los investigadores pueden entonces utilizar estas jerarquías para restringir el tipo de estudios que incluyen en su revisión sistemática de la literatura.

Recientemente, Petticrew y Roberts [25] han sugerido que esta idea es demasiado simplista. Señalan que algunos tipos de diseño son mejores que otros para abordar diferentes tipos de cuestiones. Por ejemplo, los estudios cualitativos son más apropiados que los experimentos aleatorios para evaluar si los profesionales encuentran una nueva tecnología apropiada para el tipo de aplicaciones que tienen que construir. Por lo tanto, si queremos limitarnos a estudios de un tipo específico, deberíamos restringirnos a los estudios que son más adecuados para abordar nuestras preguntas específicas de investigación.

Sin embargo, hay pruebas de que los estudios de observación (por ejemplo, de correlación) pueden ser poco fiables. Los investigadores médicos han descubierto a menudo que los resultados de los estudios de observación a muy gran escala han sido anulados por los resultados de los ensayos controlados aleatorios. Un ejemplo reciente es el de los supuestos beneficios de la vitamina C [22]. Dos estudios de observación a gran escala habían sugerido previamente que tomar vitamina C

protegía contra las enfermedades cardíacas. Lawlor y otros [22] sugieren que la razón por la que los estudios de observación encontraron un resultado que no pudo observarse en los ensayos aleatorios fue que el uso de la vitamina C era un sustituto de otras características del estilo de vida que protegen contra las enfermedades cardíacas, como el ejercicio y el mantenimiento de una dieta saludable. Esto es

una cuestión que debe ser tomada en serio en la ingeniería de software, donde gran parte de nuestras investigaciones sobre temas como la estimación de los costos de software y los factores de éxito de los proyectos son estudios de correlación. Los buenos estudios de observación deben considerar los posibles efectos de confusión, establecer métodos para medirlos y ajustar cualquier análisis para permitir su efecto. En particular, deben incluir análisis de sensibilidad para investigar el impacto de los factores de confusión medidos y no medidos.

6.3.2 Desarrollo de instrumentos de calidad

Las evaluaciones detalladas de la calidad suelen basarse en "instrumentos de calidad", que son listas de verificación de los factores que deben evaluarse para cada estudio. Si se asignan escalas numéricas a los elementos de calidad de una lista de verificación, se pueden obtener evaluaciones numéricas de la calidad.

Las listas de verificación suelen derivar de la consideración de factores que podrían sesgar los resultados de los estudios. Las Directrices del CRD [19], las Directrices del Consejo Nacional de Salud e Investigación Médica de Australia [1] y el Manual del Revisor Cochrane [7] se refieren a cuatro tipos de sesgo que se muestran en la Tabla 4. (Hemos modificado las definiciones (ligeramente) y los mecanismos de protección (considerablemente) para abordar la ingeniería de software en lugar de la medicina). En particular, los investigadores médicos dependen de sujetos y experimentadores "cegadores" (es decir, que se aseguran de que ni el sujeto ni el investigador sepan a qué tratamiento se asigna un sujeto) para abordar el sesgo de rendimiento y medición. Sin embargo, ese protocolo suele ser imposible para los experimentos de ingeniería de programas informáticos.

Cuadro 4 Tipos de sesgo

Escriba	Sinónimos	Definición	Mecanismo de protección
Sesgo de selección	Sesgo de asignación	Diferencia sistemática entre los grupos de comparación con respecto al tratamiento	La asignación al azar de un gran número de sujetos con ocultación del método de asignación (por ejemplo, la asignación mediante un programa informático no elección del experimentador).
Sesgo de rendimiento		La diferencia sistemática es la realización de grupos de comparación aparte del tratamiento que se está evaluando.	Replicación de los estudios utilizando diferentes experimentadores. El uso de experimentadores sin interés personal en cualquiera de los dos tratamientos.
Sesgo de medición	Detección de sesgos	La diferencia sistemática entre los grupos en la forma en que se determinan los resultados.	Cegando a los evaluadores de resultados a los tratamientos son a veces posibles.
Sesgo de desgaste	Sesgo de exclusión	Diferencias sistemáticas entre los grupos de comparación en lo que respecta a los retiros o exclusiones de participantes del estudio muestra.	Informar de las razones de todas las retiradas. Análisis de sensibilidad que incluya a todos los participantes excluidos.

Los factores identificados en el cuadro 4 pueden refinarse en un instrumento de calidad si se consideran:

- Elementos genéricos que se relacionan con características de diseños de estudios particulares, como diseños de encuestas, diseños experimentales y diseños de estudios cualitativos.
- Elementos específicos relacionados con el tema del examen, como el método particular de validación cruzada utilizado en un estudio de la exactitud de la predicción de la estimación de los costos.

Las listas de verificación también se elaboran teniendo en cuenta los problemas de sesgo y validez que pueden producirse en las diferentes etapas de un estudio empírico:

- Diseño
- Conducta
- Análisis
- Conclusiones.

Hay muchas listas de control de calidad publicadas para diferentes tipos de estudios empíricos. Todas las directrices médicas proporcionan listas de control destinadas a ayudar a la evaluación de la calidad realizada durante una revisión sistemática de la literatura, al igual que Fink [11] y Petticrew y Roberts [25]. Además, Crombie [10] y Greenhalgh [12] también proporcionan listas de verificación destinadas a ayudar al lector a evaluar un artículo específico. Shaddish y otros [25] analizan los diseños cuasiexperimentales y proporcionan un amplio resumen de las cuestiones de validez que los afectan. Sin embargo, cada fuente identifica un conjunto de preguntas ligeramente diferente y no hay un conjunto de preguntas estándar acordado.

Para los estudios cuantitativos hemos acumulado una lista de preguntas de [10], [11], [12],
19] y 25] y los organizó con respecto a la etapa de estudio y el tipo de estudio (ver Tabla 5). No sugerimos que nadie utilice todas las preguntas. Los investigadores deberían adoptar la sugerencia de Fink [11] que consiste en revisar la lista de preguntas en el contexto de su propio estudio y seleccionar aquellas preguntas de evaluación de calidad que sean más apropiadas para sus preguntas de investigación específicas. Puede que necesiten construir una escala de medición para cada ítem, ya que a veces una simple respuesta de Sí/No puede ser engañosa. Cualquiera que sea la forma que adopte el instrumento de calidad, deberá evaluarse su fiabilidad y utilidad durante los ensayos del protocolo de estudio antes de aplicarlo a todos los estudios seleccionados.

Ejemplos

Kitchenham y otros [21] construyeron un cuestionario de calidad basado en 5 cuestiones que afectaban a la calidad del estudio y que se puntuaron para proporcionar una medida general de la calidad del estudio:

1. ¿Es apropiado el proceso de análisis de datos?
   - 1.1 ¿Se investigaron los datos para identificar los valores atípicos y evaluar las propiedades de distribución antes del análisis?
   - 1.2 ¿Se utilizó el resultado de la investigación de manera adecuada para transformar los datos y seleccionar los puntos de datos apropiados?
2. ¿Se han realizado estudios de sensibilidad o análisis residuales?
   - 2.1 ¿Estuvieron los modelos de estimación resultantes sujetos a un análisis de sensibilidad o residual?
   - 2.2 ¿Se utilizó el resultado del análisis de sensibilidad o residual para eliminar los puntos de datos anormales en caso necesario?
3. ¿Se basaron las estadísticas de precisión en la escala de datos en bruto?
4. ¿Qué tan bueno fue el método de comparación del estudio?
   - 4.1 ¿Fue la única compañía seleccionada al azar (no seleccionada por conveniencia) de varias compañías diferentes?
   - 4.2 ¿La comparación se basó en una muestra independiente de resistir (0,5) o en subconjuntos aleatorios (0,33), sin resistir (0,17), sin resistir (0)? Las puntuaciones utilizadas para este punto reflejan la opinión de los investigadores con respecto a la rigurosidad de cada criterio.
5. El tamaño del conjunto de datos internos, medido según los criterios que se presentan a continuación. Siempre que un estudio utilizó más de un conjunto de datos dentro de la

empresa, se utilizó la puntuación media:

- Menos de 10 proyectos: Mala calidad (puntuación = 0)
- Entre 10 y 20 proyectos: Calidad aceptable (puntuación = 0,33)
- Entre 21 y 40 proyectos: Buena calidad (puntuación = 0,67)
- Más de 40 proyectos: Excelente calidad (puntuación = 1)

También examinaron la calidad de la información sobre la base de cuatro preguntas:

1. ¿Está claro qué proyectos se utilizaron para construir cada modelo?
2. ¿Está claro cómo se midió la precisión?
3. ¿Está claro qué método de validación cruzada se utilizó?
4. ¿Se definieron plenamente todos los métodos de construcción de modelos (instrumentos y métodos utilizados)?

Es una buena práctica no incluir la calidad del estudio y la calidad de las puntuaciones de los informes en una sola medida, pero Kitchenham y otros propusieron utilizar una medida ponderada que diera menos peso a la puntuación de calidad de los informes.

El cuestionario de calidad de Kitchenham y otros se basó en la naturaleza específica de los estudios primarios (como el método de validación cruzada utilizado), así como en cuestiones de calidad más generales (como el tamaño de la muestra y el análisis de sensibilidad).

Jørgensen [17] no realizó una evaluación específica de la calidad de los estudios primarios.

Cuadro 5 Lista resumida de verificación de la calidad de los estudios cuantitativos

Pregunta	Estudios Empíricos Cuantitativos (no hay un tipo específico)	La correlación (observacional estudios)	Encuestas	Experimentos	Fuente
Diseño					
¿Están los objetivos claramente establecidos?	X	X	X	X	[11], [10]
¿El estudio fue diseñado con estas preguntas en mente?			X		[25]
¿Permiten las medidas de estudio responder a las preguntas?			X	X	[10], [25]
¿Qué población se estaba estudiando?			X		[25]
¿Quién estaba incluido?			X		[12]
¿Quién fue excluido?			X		[12]
¿Cómo se obtuvo la muestra (por ejemplo, postal, entrevista, web-basado en)?			X		[10], [12], [25]
¿Es probable que el método de encuesta haya introducido un sesgo significativo?			X		[25]
¿Es la muestra representativa de la población a la que el los resultados se generalizarán?			X	X	[10], [25]
¿Se asignaron los tratamientos al azar?				X	[10]
¿Hay un grupo de comparación o de control?	X		X	X	[12]
Si hay un grupo de control, ¿son los participantes similares a los de los participantes en el grupo de tratamiento en términos de variables que pueden afectar a los resultados del estudio?	X		X	X	[10], [12]
¿Se justificó el tamaño de la muestra	X		X	X	[10], [12]
Si el estudio implica la evaluación de una tecnología, es la tecnología claramente definida?	X	X	X	X	[11]
¿Podría la elección de los sujetos influir en el tamaño del efecto del tratamiento?				X	[10], [11], [19],[25]
¿Podría la falta de ceguera introducir un sesgo?				X	[10]
¿Se miden adecuadamente las variables utilizadas en el estudio (es decir, ¿es probable que las variables sean válidas y	X	X	X	X	[10], [11], [19],[25]

fiables)?					
¿Están plenamente definidas las medidas utilizadas en el estudio?	X	X	X	X	[11]

¿Son las medidas utilizadas en el estudio las más relevantes para responder a las preguntas de la investigación?	X	X	X	X	[11], [19],[25]
¿Es el alcance (tamaño y longitud) del estudio suficiente para permiten identificar los cambios en los resultados de interés?	X		X	X	[19], [12], [25]
Conducta					
¿Ocurrieron eventos adversos durante el estudio?	X	X	X	X	[10]
¿La evaluación de resultados fue ciega al grupo de tratamiento?	X			X	[19], [12], [25]
¿Se describen adecuadamente los métodos de reunión de datos?	X	X	X	X	[11]
Si se comparan dos grupos, ¿se trataron de manera similar dentro del estudio?				X	[12], [25]
Si el estudio involucra a los participantes a lo largo del tiempo, ¿qué proporción de las personas que se inscribieron al principio abandonaron?	X		X	X	[10], [11]
¿Cómo se llevó a cabo la aleatorización?				X	[10]
Análisis					
¿Cuál fue la tasa de respuesta?			X		[10], [25]
¿Se informó el denominador (es decir, el tamaño de la población)?			X		[25]
¿Explican los investigadores los tipos de datos (continuos, ordinales, categóricos)?	X	X	X	X	[11]
¿Están los participantes del estudio o las unidades de observación adecuadamente descritos? Por ejemplo, la experiencia en el SE, el tipo (estudiante, practicante, consultor), la nacionalidad, la experiencia en la tarea y otras variables relevantes.	X	X	X	X	[12], [25]
¿Se describieron adecuadamente los datos básicos?	X	X	X	X	[10]
¿Los "abandonos" han introducido un sesgo?	X		X	X	[11], [12], [25]
¿Se dan razones para negarse a participar?	X		X	X	[11]
¿Se describen los métodos estadísticos?	X	X	X	X	[10], [11], [19]
¿El programa estadístico utilizado para analizar los datos ¿Referenciado?	X	X	X	X	[11]
¿Están justificados los métodos estadísticos?	X	X	X	X	[11]
¿Está claro el propósito del análisis?	X	X	X	X	[11]
¿Se describen los sistemas de puntuación?	X			X	[11]

¿Se controlan adecuadamente los posibles factores de confusión en el análisis?	X	X	X	X	[11]
¿Los números se suman a través de diferentes tablas y	X	X	X	X	[10], [11]

subgrupos?					
Si los diferentes grupos eran diferentes al comienzo del estudio o fueron tratados de manera diferente durante el estudio, ¿se hizo algún intento para controlar estas diferencias, ya sea estadísticamente o por comparación?	X		X	X	[12], [25]
Si es así, ¿fue un éxito?	X		X	X	[25]
¿Se evaluó la importancia estadística?	X	X	X	X	[10]
Si se utilizan pruebas estadísticas para determinar las diferencias, es la valor p real dado?	X	X	X	X	[11]
Si el estudio se refiere a las diferencias entre grupos, ¿se dan límites de confianza que describan la magnitud de cualquier diferencia observada?	X		X	X	[11]
¿Existen pruebas de pruebas estadísticas múltiples o grandes números de análisis post hoc?	X	X	X	X	[10], [25]
¿Cómo podría surgir un sesgo de selección?	X		X	X	[10], [25]
¿Se informó de los efectos secundarios?					[10]
Conclusiones					
¿Todas las preguntas del estudio tienen respuesta?	X	X	X	X	[11]
¿Qué significan los principales hallazgos?	X	X	X	X	[10]
¿Se presentan los resultados negativos?	X	X	X	X	[11]
Si se utilizan pruebas estadísticas para determinar las diferencias, ¿se discute la importancia práctica?	X	X	X	X	[11]
Si los abandonos difieren de los participantes, son limitaciones a la ¿Resultados discutidos?	X		X	X	[11]
¿Cómo se interpretan los hallazgos nulos? (Es decir, tiene la posibilidad que el tamaño de la muestra es demasiado pequeño?)	X	X	X	X	[10], [12]
¿Se pasan por alto los efectos importantes?	X	X	X	X	[10]
¿Cómo se comparan los resultados con los informes anteriores?	X	X	X	X	[10]
¿Cómo se añaden los resultados a la literatura?	X	X	X	X	[12]
¿Qué consecuencias tiene el informe para la práctica?	X	X	X	X	[10]
¿Explican los investigadores las consecuencias de cualquier problema con la validez/fiabilidad de sus medidas?	X	X	X	X	[11]

Si una revisión incluye estudios cualitativos, será necesario evaluar su calidad. En el cuadro 6 figura una lista de verificación para evaluar la calidad de los estudios cualitativos.

Cuadro 6 Lista de verificación de los estudios cualitativos

Número	Pregunta	Fuente
1	¿Qué tan creíbles son los hallazgos?	[12], [25]
1.1	Si son creíbles, ¿son importantes?	[12]
2	¿Cómo se ha extendido el conocimiento o la comprensión por parte de la ¿Investigación?	[12], [25]
3	¿Qué tan bien aborda la evaluación sus objetivos y propósitos originales?	[25]
4	¿Qué tan bien se explica el alcance de la inferencia más amplia?	[25]
5	¿Qué tan clara es la base de la evaluación de la evaluación?	[25]
6	¿Qué tan defendible es el diseño de la investigación?	[12], [25], [11]
7	¿Qué tan bien definidos están el diseño de la muestra y la selección del objetivo de casos/documentos?	[12], [25], [11]
8	¿Qué tan bien se describe la composición y cobertura de la muestra eventual?	[25]
9	¿Qué tan bien se llevó a cabo la recolección de datos?	[12], [25], [11]
10	¿Qué tan bien ha funcionado el enfoque y la formulación del análisis ...se ha transmitido?	[12], [25], [11]
11	¿Qué tan bien se conservan los contextos y las fuentes de datos y retratado?	[25]
12	¿Hasta qué punto se ha explorado la diversidad de perspectivas y contextos?	[25]
13	¿Qué tan bien tienen los detalles, la profundidad y la complejidad (es decir, la riqueza) de los datos que se han transmitido?	[25]
14	¿Qué tan claros son los vínculos entre los datos, la interpretación y conclusiones - es decir, ¿qué tan bien se puede ver el camino hacia cualquier conclusión?	[25]
15	¿Qué tan clara y coherente es la información?	[25]
16	¿Qué tan claras son las suposiciones, las perspectivas teóricas y los valores que han dado forma a la forma y el resultado de la evaluación?	[12], [25], [11]
17	¿Qué pruebas hay de que se presta atención a las cuestiones éticas?	[25]
18	¿Qué tan adecuadamente se ha documentado el proceso de investigación?	[25]

6.3.3 Usando el instrumento de calidad

Es importante que los investigadores no sólo definan el instrumento de calidad en el protocolo de estudio, sino que también especifiquen cómo se utilizarán los datos de calidad. Los datos de calidad pueden utilizarse de dos maneras bastante diferentes:

1. Para ayudar a la selección del estudio primario. En este caso, los datos de calidad se utilizan para construir criterios detallados de inclusión/exclusión. Los datos de calidad deben recopilarse antes de la actividad principal de reunión de datos utilizando formularios de reunión de datos separados.
2. Para ayudar al análisis y síntesis de los datos. En este caso los datos de calidad se utilizan para identificar subconjuntos del estudio primario para investigar si las

diferencias de calidad están asociadas a diferentes resultados del estudio primario. Los datos de calidad pueden recogerse al mismo tiempo que la actividad principal de extracción de datos utilizando un formulario conjunto.

Por supuesto, es posible tener ambos tipos de datos de calidad en el mismo examen sistemático.

Ejemplo

Kitchenham y otros [21] usaron la puntuación de calidad para investigar si los resultados del estudio primario estaban asociados con la calidad del estudio. También investigaron si algunos de los factores de calidad individuales (es decir, el tamaño de la muestra, el método de validación) estaban asociados con el resultado del estudio primario.

Algunos investigadores han sugerido ponderar los resultados de los meta-análisis utilizando puntuaciones de calidad. Esta idea **no está recomendada** por ninguna de las directrices médicas.

Si un examen sistemático incluye estudios de diferentes tipos, es necesario utilizar un instrumento de calidad apropiado para cada tipo de estudio. En algunos casos, un conjunto común de preguntas de evaluación de la calidad puede ser adecuado para todos los estudios cuantitativos incluidos en un examen sistemático, pero si un examen incluye estudios cualitativos y cuantitativos será esencial utilizar diferentes listas de verificación.

6.3.4 Limitaciones de la evaluación de la calidad

Los estudios primarios suelen estar mal informados, por lo que tal vez no sea posible determinar cómo evaluar un criterio de calidad. Es tentador asumir que porque algo no se informó, no se hizo. Esta suposición puede ser incorrecta. Los investigadores deben tratar de obtener más información de los autores del estudio. Petticrew y Roberts [25] señalan explícitamente que las listas de control de calidad deben abordar la calidad metodológica *y no la calidad de los informes*.

Hay pruebas limitadas de las relaciones entre los factores que se cree que afectan a la validez y los resultados reales de los estudios. Las pruebas indican que la ocultación inadecuada de la asignación y la falta de doble cegamiento dan lugar a sobreestimaciones de los efectos del tratamiento, pero la repercusión de otros factores de calidad no está respaldada por pruebas empíricas.

Es posible identificar un análisis estadístico inadecuado o inapropiado, pero sin acceso a los datos originales no es posible corregir el análisis. Muy a menudo los datos de los programas informáticos son confidenciales y, por lo tanto, no pueden ponerse a disposición de los investigadores en general.
En algunos casos, los ingenieros de software pueden negarse a poner sus datos a disposición de otros investigadores porque quieren seguir publicando análisis de los datos.

6.4 Extracción de datos

El objetivo de esta etapa es diseñar formularios de extracción de datos para registrar con precisión la información que los investigadores obtienen de los estudios primarios. Para reducir la posibilidad de sesgo, los formularios de extracción de datos deben definirse y probarse cuando se defina el protocolo de estudio.

6.4.1 Diseño de formularios de extracción de datos

Los formularios de extracción de datos deben diseñarse de manera que se reúna toda la información necesaria para abordar las preguntas de examen y los criterios de calidad del estudio. Si los criterios de calidad se van a utilizar para identificar los criterios de inclusión/exclusión, requieren formularios separados (ya que la información se debe recopilar antes del ejercicio principal de extracción de datos). Si los criterios de calidad han de utilizarse como parte del análisis de los datos, los criterios de calidad y los datos del examen pueden incluirse en el mismo formulario.

En la mayoría de los casos, la extracción de datos definirá un conjunto de valores numéricos que deberán extraerse para cada estudio (por ejemplo, número de sujetos, efecto del tratamiento, intervalos de confianza, etc.). Los datos numéricos son importantes para cualquier intento de resumir los resultados

de un conjunto de estudios primarios y son un requisito previo para el metaanálisis (es decir, técnicas estadísticas destinadas a integrar los resultados de los estudios primarios).

Los formularios de extracción de datos deben ser probados en una muestra de estudios primarios. Si varios investigadores utilizan los formularios, todos deben participar en el piloto. Los estudios piloto tienen por objeto evaluar tanto las cuestiones técnicas, como la integridad de los formularios, como las cuestiones de utilidad, como la claridad de las instrucciones para el usuario y el orden de las preguntas.

Los formularios electrónicos son útiles y pueden facilitar el análisis posterior.

6.4.2 Contenido de los formularios de recopilación de datos

Además de incluir todas las preguntas necesarias para responder a la pregunta de examen y los criterios de evaluación de la calidad, los formularios de reunión de datos deben proporcionar información estándar, entre otras cosas:

- Nombre del revisor
- Fecha de la extracción de datos
- Título, autores, revista, detalles de la publicación
- Espacio para notas adicionales

Ejemplos

Kitchenham y otros [21] usaron el formulario de extracción que se muestra en la Tabla 7 (nótese que el formulario real también incluía las preguntas de calidad).

Cuadro 7 Formulario de recopilación de datos completado para Maxwell y otros, 1998

Datos	Valor	Notas adicionales
Extractor de datos		
Comprobador de datos		
Identificador del estudio	S1	
Dominio de la aplicación	Espacio, militar e industrial	
Nombre de la base de datos	Agencia Espacial Europea (ESA)	
Número de proyectos en la base de datos (incluidos los de proyectos de empresa)	108	
Número de proyectos interempresariales	60	
Número de proyectos en el conjunto de datos internos de la empresa	29	
Métrica de tamaño: FP (Sí/No) Versión utilizada: LOC (Sí/No) Versión utilizada: Otros (Sí/No) Número:	FP: No LOC: Sí (KLOC) Otros: No	
Número de empresas	37	
Número de países representados	8	Sólo europeo.
¿Se aplicaron controles de calidad a la reunión de datos?	No	

Si el control de calidad, por favor describa		
¿Cómo se midió la precisión?	Medidas: R2 (sólo para la construcción del modelo) MMRE Pred(25) r (Correlación entre la estimación y el real)	

Modelo de compañía cruzada		
¿Qué técnica(s) se utilizó para construir el modelo de compañía cruzada?	Se utilizó un análisis preliminar de la productividad para identificar los factores que debían incluirse en el modelo de estimación del esfuerzo. Modelos lineales generalizados (usando SAS). Se investigaron los modelos multiplicativo y aditivo. El modelo multiplicativo es un modelo logarítmico.	
Si se utilizaron varias técnicas, ¿cuál fue la más precisa?	En todos los casos, la evaluación de la exactitud se basó en los modelos logarítmicos y no en los modelos aditivos.	Se puede asumir que los modelos lineales no funcionaron bien.
¿Qué transformaciones se utilizaron, si es que se utilizaron alguna?	No está claro si las variables fueron transformadas o si el GLM fue usado para construir un modelo logarítmico lineal	No es importante: se utilizaron los modelos de registro y se presentaron en la forma de datos en bruto - por lo tanto cualquier métrica de precisión fue basado en las predicciones de los datos en bruto.
¿Qué variables se incluyeron en el modelo de compañía cruzada?	KLOC, subconjunto de idiomas, subconjunto de categorías, RELY	La categoría es el tipo de aplicación. RELY es la fiabilidad definida por Boehm (1981)
¿Qué método de validación cruzada se utilizó?	Una muestra de 9 proyectos de la única empresa se utilizó para evaluar la precisión de la estimación	
¿Se comparó el modelo de compañía cruzada con una línea de base para comprobar si era mejor que la oportunidad?	Sí	La línea de base fue la correlación entre las estimaciones y los datos reales para el atraco.
¿Qué medida(s) se utilizó(aron) como punto de referencia?	La correlación entre la predicción y el real para la empresa individual se probó para determinar su importancia estadística. (Nótese que fue significativamente diferente de cero para el conjunto de datos de 20 proyectos, pero no el conjunto de datos de 9 proyectos de retención).	
Modelo dentro de la empresa		
¿Qué técnica(s) se utilizó para construir el modelo interno de la empresa?	Se utilizó un análisis preliminar de la productividad para identificar los factores que debían incluirse en el modelo de estimación del esfuerzo. Modelos lineales generalizados (usando SAS). Se investigaron los modelos multiplicativo y aditivo. El El modelo multiplicativo es un modelo logarítmico.	

Si varias técnicas fueran usó la que era más precisa?	En todos los casos, la evaluación de la exactitud se basaba en los modelos logarítmicos no en los modelos aditivos.	Se puede suponer que los modelos lineales no funcionaron bien.
¿Qué transformaciones se utilizaron, si es que se utilizaron alguna?	No está claro si las variables fueron transformadas o si el GLM fue usado para construir un modelo logarítmico lineal	No es importante: se utilizaron los modelos logísticos y se presentaron en forma de datos en bruto, por lo que cualquier métrica de precisión se basó en datos en bruto predicciones.
¿Qué variables se incluyeron en el interior... modelo de empresa?	KLOC, subconjunto de idiomas, Año	
¿Qué validación cruzada	Una muestra de 9 proyectos	

se utilizó el método	de la única empresa se utilizó para evaluar la precisión de la estimación	
Comparación		
¿Cuál fue la precisión obtenida usando el modelo de compañía cruzada?	Exactitud en el principal conjunto de datos de una sola empresa (modelo de registro): n=11 (9 proyectos omitidos) MMRE=50% Pred(25)=27% r=0.83 La precisión en el conjunto de datos de una sola empresa de retención n=4 (5 proyectos omitidos) MMRE=36% Pred(25)=25% R=0.16 (n.s)	Utilizando los 79 proyectos de empresas cruzadas, Maxwell y otros identificaron el mejor modelo para ese conjunto de datos y el mejor modelo para los datos de una sola empresa. Los dos modelos eran idénticos. Estos datos indican que para todos los proyectos de una sola compañía: n=15 Pred(25)=26,7% (4 de 15) MMRE=46,3%
¿Cuál fue la precisión obtenida usando el modelo interno de la compañía?	Exactitud en el principal conjunto de datos de una sola empresa (modelo de registro): n=14 (6 proyectos omitidos) R2=0,92 MMRE=41% Pred(25)=36% r=0.99 La precisión en el conjunto de datos de una sola empresa de retención n=6 (3 proyectos omitidos) MMRE=65% Pred(25)=50% (3 de 6) r=0.96	
¿Qué medida se utilizó para comprobar la importancia estadística de la exactitud de la predicción (por ejemplo, residuos absolutos, MREs)?	Esfuerzo estimado y real	
¿Qué pruebas estadísticas se utilizaron para comparar los resultados?	r, la correlación entre la predicción y la realidad	
¿Cuáles fueron los resultados de las pruebas?		
Resumen de datos		
Resumen de la base de datos (todos los proyectos) para la medición del tamaño y el esfuerzo.	Esfuerzo mínimo: 7,8 MM Esfuerzo máximo: 4361 MM Esfuerzo medio: 284 MM Esfuerzo mediano: 93 MM Tamaño min: 2000 KLOC Tamaño máximo: 413000 KLOC Tamaño medio: 51010 KLOC Tamaño medio: 22300 KLOC	KLOC: no en blanco, sin comentarios entregó 1000 líneas. Para el código reutilizado se hicieron los ajustes de Boehm (Boehm, 1981). El esfuerzo se midió en meses-hombre, con 144 horas-hombre por mes-hombre

Resumen de los datos de la empresa para la medición del tamaño y el esfuerzo.	Esfuerzo mínimo: Esfuerzo máximo: Esfuerzo medio: Esfuerzo mediano: Tamaño min: Tamaño max: Tamaño medio: La mediana del tamaño:	No se especifica

Jørgensen [17] extrajo los factores de diseño y los resultados del estudio primario. Se incluyeron los factores de diseño:

- Diseño del estudio
- Proceso de selección del método de estimación
- Modelos de estimación

- Nivel de calibración
- Experiencia en el uso de modelos y grado de utilización mecánica del modelo
- El proceso de juicio de los expertos
- Estimación del juicio de los expertos
- Posibles sesgos motivacionales en la situación de estimación
- Aportación de la estimación
- Información contextual
- Complejidad de la estimación
- Limitaciones de la equidad
- Otras cuestiones de

diseño Los resultados del
estudio incluyeron

- Precisión
- Variación
- Otros resultados

El artículo de Jørgensen incluye el formulario de extracción completado para cada estudio primario.

6.4.3 Procedimientos de extracción de datos

Siempre que sea posible, la extracción de datos debe ser realizada independientemente por dos o más investigadores. Los datos de los investigadores deben compararse y los desacuerdos deben resolverse ya sea por consenso entre los investigadores o por arbitraje de un investigador independiente adicional. Las incertidumbres acerca de cualquier fuente primaria sobre la que no se pueda llegar a un acuerdo deben investigarse como parte de cualquier análisis de sensibilidad. Debe utilizarse un formulario separado para marcar y corregir errores o desacuerdos.

Si varios investigadores examinan cada uno diferentes estudios primarios porque las limitaciones de tiempo o de recursos impiden que todos los documentos primarios sean evaluados por dos investigadores como mínimo, es importante emplear algún método para comprobar que los investigadores extraen los datos de manera coherente. Por ejemplo, algunos trabajos deben ser examinados por todos los investigadores (por ejemplo, una muestra aleatoria de estudios primarios), de modo que pueda evaluarse la coherencia entre los investigadores.

Para los investigadores solteros, como los estudiantes de doctorado, se deben utilizar otras técnicas de comprobación. Por ejemplo, los supervisores podrían realizar la extracción de datos en una muestra aleatoria de los estudios primarios y sus resultados se cotejarán con los del estudiante.
Como alternativa, se puede utilizar un proceso de test-retest en el que el investigador realiza una segunda extracción de una selección aleatoria de estudios primarios para comprobar la consistencia de la extracción de datos.

Ejemplos
Kitchenham y otros [21] asignaron a una persona para que fuera el extractor de datos que completó el formulario de extracción de datos y a otra persona para que fuera el verificador de datos que confirmó que los datos del formulario de extracción eran correctos. Debido a que Kitchenham y Mendes fueron coautores de algunos de los estudios primarios, también se aseguraron de que el extractor de datos nunca fuera un coautor del estudio primario. Cualquier desacuerdo era examinado y se registraba un valor de datos final acordado.

Como investigador único, Jørgensen [17] extrajo todos los datos él mismo. Sin embargo, envió los datos de cada estudio primario a un autor del estudio y pidió que le informaran si alguno de los datos extraídos era incorrecto.

6.4.4 Múltiples publicaciones de los mismos datos

Es importante no incluir múltiples publicaciones de los mismos datos en una síntesis de examen sistemático porque los informes duplicados sesgarían gravemente los resultados. Puede ser necesario ponerse en contacto con los autores para confirmar si los informes se refieren al mismo

estudio. Cuando hay publicaciones duplicadas, se debe utilizar la más completa. Incluso puede ser necesario consultar todas las versiones del informe para obtener todos los datos necesarios.

6.4.5 Datos no publicados, datos que faltan y datos que requieren manipulación

Si se dispone de información de estudios en curso, debe incluirse, siempre que se pueda obtener información de calidad apropiada sobre el estudio y se cuente con el permiso escrito de los investigadores.

Los informes no siempre incluyen todos los datos pertinentes. También pueden estar mal escritos y ser ambiguos. Una vez más, se debe contactar a los autores para obtener la información requerida.

A veces los estudios primarios no proporcionan todos los datos, pero es posible recrear los datos requeridos manipulando los datos publicados. Si se requiere alguna manipulación de ese tipo, los datos deben notificarse primero en la forma en que se publicaron. Los datos obtenidos mediante manipulación deben ser objeto de un análisis de sensibilidad.

6.4.6 Lecciones aprendidas sobre la extracción de datos

Brereton y otros [5] identificaron dos cuestiones de importancia durante la extracción de datos:

- El hecho de que un lector actúe como extractor de datos y otro como verificador de datos puede ser útil cuando hay un gran número de documentos que revisar.
- Los miembros del equipo de revisión deben asegurarse de que entienden el protocolo y el proceso de extracción de datos.

6.5 Síntesis de datos

La síntesis de los datos implica cotejar y resumir los resultados de los estudios primarios incluidos. La síntesis puede ser descriptiva (no cuantitativa). Sin embargo, a veces es posible complementar una síntesis descriptiva con un resumen cuantitativo. El uso de técnicas estadísticas para obtener una síntesis cuantitativa se denomina *metaanálisis*. La descripción de los métodos de metaanálisis está fuera del alcance de este documento, aunque se describirán las técnicas para mostrar los resultados cuantitativos. (Para saber más sobre el meta-análisis véase [7].)

Las actividades de síntesis de datos deben especificarse en el protocolo de examen. Sin embargo, algunas cuestiones no pueden resolverse hasta que se analicen realmente los datos; por ejemplo, no se requiere un análisis de subconjuntos para investigar la heterogeneidad si los resultados no muestran pruebas de heterogeneidad.

6.5.1 Síntesis descriptiva (narrativa)

La información extraída de los estudios (es decir, intervención, población, contexto, tamaño de la muestra, resultados, calidad del estudio) debe tabularse de manera coherente con la pregunta de revisión. Las tablas deben estructurarse de manera que se destaquen las similitudes y diferencias entre los resultados de los estudios.

Es importante determinar si los resultados de los estudios son coherentes entre sí (es decir, homogéneos) o incoherentes (por ejemplo, heterogéneos). Los resultados

pueden tabularse para mostrar el impacto de las posibles fuentes de heterogeneidad, por ejemplo, el tipo de estudio, la calidad del estudio y el tamaño de la muestra.

Ejemplos
Kitchenham y otros [21] tabularon los datos de los estudios primarios en tres tablas separadas basadas en el resultado del estudio primario: ninguna diferencia significativa entre el modelo entre compañías y el modelo dentro de la compañía, el modelo dentro de la compañía significativamente mejor que el modelo entre compañías y no se realizaron pruebas estadísticas. También destacaron los estudios que consideraban que debían excluirse de la síntesis porque eran réplicas completas en cuanto a la base de datos entre empresas y la base de datos dentro de la empresa porque no ofrecían pruebas adicionales independientes.

Llegaron a la conclusión de que las pequeñas empresas que producen software especializado (nicho) no se beneficiarían de la utilización de un modelo de estimación entre empresas. Las grandes empresas que producen aplicaciones de tamaño similar al de los proyectos interempresariales podrían encontrar útiles los modelos interempresariales.

Jørgensen [17] tabuló los estudios de acuerdo con la precisión relativa del modelo y los expertos. Así, consideró la precisión del experto más exacto y del menos exacto en comparación con los modelos más exactos y menos exactos. También consideró la precisión media de los modelos y los expertos. Codificó los estudios cronológicamente (al igual que Kitchenham y otros), de modo que fue posible buscar posibles asociaciones con la edad del estudio y el resultado.

Llegó a la conclusión de que los modelos no son sistemáticamente mejores que los expertos para la estimación de los costos de los programas informáticos, posiblemente porque los expertos poseen más información que los modelos o puede ser difícil construir modelos precisos de estimación del desarrollo de programas informáticos. Es probable que la opinión de los expertos sea útil si los modelos no están calibrados para la empresa que los utiliza y/o si los expertos tienen acceso a información contextual importante que puedan explotar. Los modelos (o una combinación de modelos y expertos) pueden ser útiles cuando hay sesgos situacionales hacia el exceso de optimismo, los expertos no tienen acceso a grandes cantidades de información contextual y/o los modelos están calibrados para el entorno.

6.5.2 Síntesis cuantitativa

Los datos cuantitativos también deben presentarse en forma tabular, incluyendo:

- Tamaño de la muestra para cada intervención.
- Estima el tamaño del efecto para cada intervención con errores estándar para cada efecto.
- Diferencia entre los valores medios de cada intervención y el intervalo de confianza de la diferencia.
- Unidades utilizadas para medir el efecto.

Sin embargo, para sintetizar los resultados cuantitativos de los diferentes estudios, los resultados de los estudios deben presentarse de manera comparable. Las directrices médicas sugieren diferentes medidas del efecto para diferentes tipos de resultados.

Los resultados binarios (Sí/No, Éxito/Fracaso) pueden medirse de varias maneras diferentes:

- Las probabilidades. La relación entre el número de sujetos en un grupo con un evento y el número sin un evento. Así, si 20 proyectos en un grupo de 100 proyectos no logran alcanzar los objetivos presupuestarios, las probabilidades serían de 20/80 o 0,25.
- Riesgo (proporción, probabilidad, tasa) La proporción de sujetos de un grupo en el que se ha observado un suceso. Así, si 20 de cada 100 proyectos no logran alcanzar las metas presupuestarias, el riesgo sería de 20/100 ó 0,20.

- Ratio de probabilidades (OR). La relación entre las probabilidades de un evento en el grupo experimental (o de intervención) y las probabilidades de un evento en el grupo de control. Una OR igual a uno indica que no hay diferencia entre el grupo de control y el de intervención. En el caso de los resultados no deseados, un valor inferior a uno indica que la intervención fue

éxito en la reducción del riesgo, para un resultado deseable un valor mayor que uno indica que la intervención tuvo éxito en la reducción del riesgo.

- Riesgo relativo (RR) (relación de riesgo, relación de tasas). Es la relación entre el riesgo en el grupo de intervención y el riesgo en el grupo de control. Un RR de uno indica que no hay diferencia entre los grupos de comparación. Para los eventos indeseables un RR menor que uno indica que la intervención tuvo éxito, para los eventos deseables un RR mayor que uno indica que la intervención tuvo éxito.
- Reducción del riesgo absoluto (ARR) (diferencia de riesgo, diferencia de tasas). La diferencia absoluta en la tasa de eventos entre los grupos de comparación. Una diferencia de cero indica que no hay diferencia entre los grupos. Para un resultado no deseado, una ARR menor que cero indica una intervención exitosa, para un resultado deseable una ARR mayor que cero indica una intervención exitosa.

Cada una de estas medidas tiene ventajas y desventajas. Por ejemplo, las probabilidades y los porcentajes de probabilidad son criticados por no ser bien comprendidos por quienes no son estadísticos (que no sean jugadores), mientras que las medidas de riesgo son generalmente más fáciles de comprender. Por otra parte, los estadísticos prefieren los odds ratios porque tienen algunas propiedades matemáticamente deseables. Otra cuestión es que las medidas relativas suelen ser más coherentes que las absolutas para el análisis estadístico, pero los encargados de adoptar decisiones necesitan valores absolutos para evaluar el beneficio real de una intervención.

Las medidas del efecto de los datos continuos incluyen:

- Diferencia media. La diferencia entre las medias de cada grupo (grupo de control e intervención).
- Diferencia media ponderada (WMD). Cuando los estudios han medido la diferencia en la misma escala, el peso dado a cada estudio suele ser el inverso de la varianza del estudio
- Diferencia media estandarizada (SMD). Un problema común al resumir los resultados es que éstos se suelen medir de diferentes maneras, por ejemplo, la productividad se puede medir en puntos de función por hora, o líneas de código por día. La calidad podría medirse como la probabilidad de presentar uno o más fallos o el número de fallos observados. Cuando los estudios utilizan diferentes escalas, la diferencia media puede dividirse por una estimación de la desviación estándar dentro de los grupos para producir un valor normalizado sin ninguna unidad. Sin embargo, las DME sólo son válidas si la diferencia en las desviaciones estándar refleja diferencias en la escala de medición, no diferencias reales entre las poblaciones de ensayo.

6.5.3 Presentación de los resultados cuantitativos

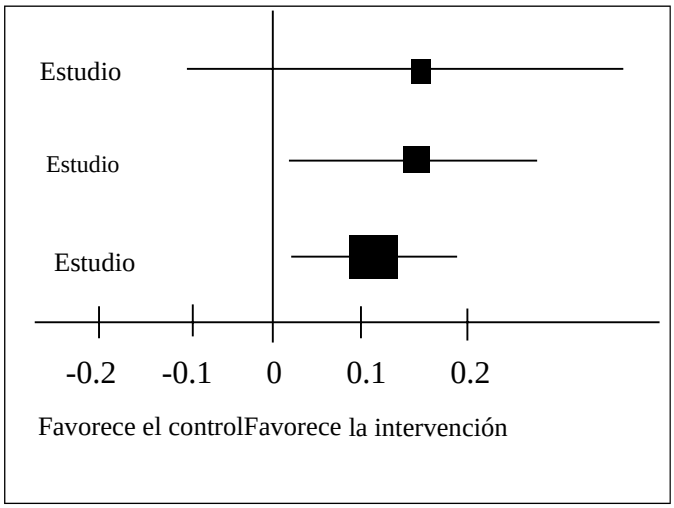
El mecanismo más común para presentar los resultados cuantitativos es un diagrama de bosque, como se muestra en la figura 1. Una parcela forestal presenta la media y la varianza de la diferencia para cada estudio. La línea representa el error estándar de la diferencia, el recuadro representa la diferencia media y su tamaño es proporcional al número de sujetos del estudio. Una parcela forestal también puede ser anotada con la información numérica que indica el número de sujetos en cada grupo, la diferencia media y el intervalo de confianza en la media. Si se realiza un metaanálisis formal, la entrada inferior en una parcela forestal será la estimación resumida de la diferencia de tratamiento y el intervalo de confianza de la diferencia resumida.

La figura 1 representa el resultado ideal de un resumen cuantitativo, ya que los resultados de los estudios coinciden básicamente. Es evidente que existe un efecto de tratamiento genuino y un único efecto global

La estadística resumida sería una buena estimación de ese efecto. Si los efectos fueran muy diferentes de un estudio a otro, nuestros resultados sugerirían heterogeneidad. Una sola estadística resumida global probablemente sería de poco valor. La revisión sistemática debería continuar con una investigación de las razones de la heterogeneidad.

Para evitar los problemas del análisis post hoc (es decir, la "pesca" de los resultados), los investigadores deben identificar las posibles fuentes de heterogeneidad cuando elaboren el protocolo de examen. Por ejemplo, los estudios de diferentes tipos pueden tener resultados diferentes, por lo que suele ser útil sintetizar los resultados de diferentes tipos de estudios por separado y evaluar si los resultados son coherentes entre los diferentes tipos de estudios.

Figura 1 Ejemplo de una parcela forestal



6.5.4 Síntesis cualitativa

Sintetizar los estudios cualitativos supone tratar de integrar estudios que comprendan los resultados y conclusiones del lenguaje natural, en los que diferentes investigadores pueden haber utilizado términos y conceptos con significados sutilmente (o groseramente) diferentes. Noblit y Hare [23] proponen tres enfoques para la síntesis cualitativa:

- Traducción recíproca. Cuando los estudios tratan de cosas similares y los investigadores intentan proporcionar un resumen aditivo, la síntesis puede lograrse "traduciendo" cada caso a cada uno de los otros casos.
- Síntesis de refutar. Cuando los estudios son refutaciones implícitas o explícitas de cada uno de ellos, es necesario traducir tanto los estudios individuales como las refutaciones para poder analizarlas en detalle.
- Line of argument synthesis. This approach is used when researchers are concerned about what they can infer about a topic as a whole from a set of selective studies that look at a part of the issue. This analysis is a two part one. First the individual studies are analysed, then an attempt is made to analyse the set of studies as a whole. This is rather similar to a descriptive synthesis. Issues of importance are identified and the approach to each issue taken by each study is documented and tabulated.

6.5.5 Synthesis of qualitative and quantitative studies

When researchers have a systematic literature review that includes quantitative and qualitative studies, they should:

- Synthesise the quantitative and qualitative studies separately.
- Then attempt to integrate the qualitative and quantitative results by investigating whether the qualitative results can help explain the quantitative results. For example qualitative studies can suggest reasons why a treatment does or does not work in specific circumstances.

As yet we have no published software engineering SLRs that have combined a qualitative survey and a quantitative survey. However, Sutcliffe et al. [28] provide an example of such a study in their survey of children and healthy eating. They performed three syntheses:

1. A statistical meta-analysis of studies which attempted to increase children's consumption of fruit and vegetables.
2. A thematic qualitative synthesis of studies focused on children's views of healthy eating.
3. A "cross-study synthesis" that used the results of the qualitative synthesis to interpret the findings of the meta-analysis.

6.5.6 Sensitivity analysis

Sensitivity analysis is important whether you have undertaken a descriptive or quantitative synthesis. However, it is usually easier to perform as part of a meta-analysis (since quantitative sensitivity analysis techniques are well understood). In such cases, the results of the analysis should be repeated on various subsets of primary studies to determine whether the results are robust. The types of subsets selected would be:

- High quality primary studies only.
- Primary studies of particular types.
- Primary studies for which data extraction presented no difficulties (i.e. excluding any studies where there was some residual disagreement about the data extracted).
- The experimental method used by the primary studies.

When a formal meta-analysis is not undertaken but quantitative results have been tabulated, forest plots can be annotated to identify high quality primary studies, the studies can be presented in decreasing order of quality or in decreasing study type hierarchy order. Primary studies where there are queries about the data extracted can also be explicitly identified on the forest plot, by for example, using grey colouring for less reliable studies and black colouring for reliable studies.

When you have undertaken a descriptive synthesis, sensitivity analysis is more subjective, but you should consider what impact excluding poor quality studies or studies of a particular type would have on your conclusions.

Ejemplos

Jørgensen [17] reported the results of field studies as well as the results of all studies based on the argument that field studies would have more external validity.

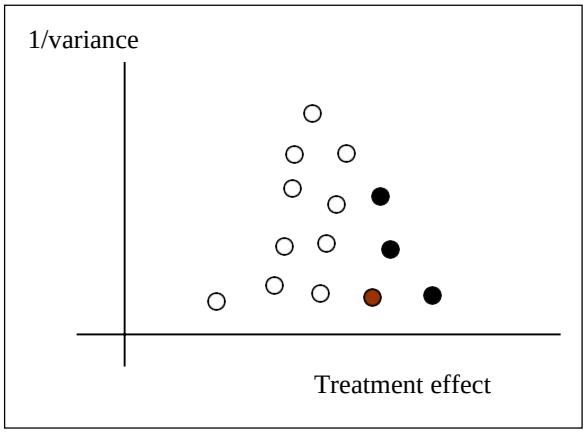
In a study of the Technology Acceptance Model (TAM), Turner et al. [29] investigated the relationship between the TAM variables Perceived Ease of Use (PEU) and PU (Perceived Usefulness) and Actual Use measured subjectively and objectively. As part of their sensitivity

analysis they investigated the impact on their results of removing primary studies authored by the researcher who developed the TAM.

6.5.7 Publication bias

Funnel plots are used to assess whether or not a systematic review is likely to be vulnerable to publication bias. Funnel plots plot the treatment effect (i.e. mean difference between intervention group and control) against the inverse of the variance or the sample size. A systematic review that exhibited the funnel shape shown in Figure 2 would be assumed **not** to be exhibiting evidence of publication bias. It would be consistent with studies based on small samples showing more variability in outcome than studies based on large samples. If, however, the points shown as filled-in black dots were not present, the plot would be asymmetric and it would suggest the presence of publication bias. This would suggest the results of the systematic review must be treated with caution.

Figure 2 An example of a funnel plot



6.5.8 Lessons Learned about Data Synthesis

Brereton et al. [5] identified three issues of importance during data extraction:

- IT and software engineering systematic reviews are likely to be qualitative (i.e. descriptive) in nature.
- Even when collecting quantitative information it may not be possible to perform meta-analysis of IT and software engineering studies because the reporting protocols vary so much between studies.
- Tabulating the data is a useful means of aggregation but it is necessary to explain how the aggregated data actually answer the research questions.

7. Reporting the review (Dissemination)

The final phase of a systematic review involves writing up the results of the review and circulating the results to potentially interested parties.

7.1 Specifying the Dissemination Strategy

It is important to communicate the results of a systematic review effectively. For this reason most guidelines recommend planning the dissemination strategy during the commissioning stage (if any) or when preparing the systematic review protocol.

Academics usually assume that dissemination is about reporting results in academic journals and/or conferences. However, if the results of a systematic review are intended to influence practitioners, other forms of dissemination are necessary. In particular:

1. Practitioner-oriented journals and magazines
2. Press Releases to the popular and specialist press
3. Short summary leaflets
4. Posters
5. Web pages
6. Direct communication to affected bodies.

7.2 Formatting the Main Systematic Review Report

Usually systematic reviews will be reported in at least two formats:

- In a technical report or in a section of a PhD thesis.
- In a journal or conference paper.

A journal or conference paper will normally have a size restriction. In order to ensure that readers are able to properly evaluate the rigour and validity of a systematic review, journal papers should reference a technical report or thesis that contains all the details.

The structure and contents of reports suggested in [19] are presented in Table 8. This structure is appropriate for technical reports and journals. For PhD theses, the entries marked with an asterisk are not likely to be relevant.

7.3 Evaluating Systematic Review Reports

Journal articles will be peer reviewed as a matter of course. Experts review PhD theses as part of the examination process. In contrast, technical reports are not usually subjected to any independent evaluation. However, if systematic reviews are made available on the Web so that results are made available quickly to researchers and practitioners, it is worth organising a peer review. If an expert panel were assembled to review the study protocol, the same panel would be appropriate to undertake peer review of the systematic review report, otherwise several researchers with expertise in the topic area and/or systematic review methodology should be approached to review the report.

The evaluation process can use the quality checklists for systematic literature reviews discussed in Section 5.1.

7.4 Lessons Learned about Reporting Systematic Literature Reviews

Brereton y otros [5] identificaron dos cuestiones de importancia durante la extracción de datos:

- Review teams need to keep a detailed record of decisions made throughout the review process.
- The software engineering community needs to establish mechanisms for publishing systematic literature reviews which may result in papers that are longer than those traditionally accepted by many software engineering outlets or that have appendices stored in electronic repositories.

Staples and Niazi [27] also emphasize the need to keep a record of what happens during the conduct of the review. They point out that you need to report deviations from the protocol.

With respect to publishing systematic literature reviews, the Journal of Information and Software Technology (http://www.elsevier.com/wps/find/homepage.cws_home) has expressed a willingness to publish systematic literature reviews.

.

Table 8 Structure and Contents of Reports of Systematic Reviews

Section	Subsection	Scope	Comments
Title*			The title should be short but informative. It should be based on the question being asked. In journal papers, it should indicate that the study is a systematic review.
Authorship*			When research is done collaboratively, criteria for determining both who should be credited as an author, and the order of author's names should be defined in advance. The contribution of workers not credited as authors should be noted in the Acknowledgements section.
Executive summary or Structured Abstract*	Contexto	The importance of the research questions addressed by the review.	A structured summary or abstract allows readers to assess quickly the relevance, quality and generality of a systematic review.
	Objectives	The questions addressed by the systematic review.	
	Methods	Data Sources, Study selection, Quality Assessment and Data extraction.	
	Results	Main finding including any meta-analysis results and sensitivity analyses.	
	Conclusiones	Implications for practice and future research.	
Antecedentes		Justification of the need for the review. Summary of previous reviews.	Description of the software engineering technique being investigated and its potential importance.
Review questions		Each review question should be specified.	Identify primary and secondary review questions. Note this section may be included in the background section.
Review Methods	Data sources and search strategy		This should be based on the research protocol. Any changes to the original protocol should be reported.
	Study selection		
	Study quality assessment		
	Data extraction		
	Data synthesis		
Included and excluded studies		Inclusion and exclusion criteria. List of excluded studies with rationale for exclusion.	Study inclusion and exclusion criteria can sometimes best be represented as a flow diagram because studies will be excluded at different stages in the review for different reasons.

Results	Findings	Description of primary studies. Results of any quantitative summaries Details of any meta-analysis.	Non-quantitative summaries should be provided to summarise each of the studies and presented in tabular form. Quantitative summary results should be presented in tables and graphs.
	Sensitivity analysis		
Discussion	Principal findings		These must correspond to the findings discussed in the results section.
	Strengths and Weaknesses	Strengths and weaknesses of the evidence included in the review. Relation to other reviews, particularly considering any differences in quality and results.	A discussion of the validity of the evidence considering bias in the systematic review allows a reader to assess the reliance that may be placed on the collected evidence.
	Meaning of findings	Direction and magnitude of effect observed in summarised studies. Applicability (generalisability) of the findings.	Make clear to what extent the results imply causality by discussing the level of evidence. Discuss all benefits, adverse effects and risks. Discuss variations in effects and their reasons (for example are the treatment effects larger on larger projects).
Conclusiones	Recommendations	Practical implications for software development.	What are the implications of the results for practitioners?
		Unanswered questions and implications for future research.	
Acknowledgements*		All persons who contributed to the research but did not fulfil authorship criteria.	
Conflict of Interest			Any secondary interest on the part of the researchers (e.g. a financial interest in the technology being evaluated) should be declared.
References and Appendices			Appendices can be used to list studies included and excluded from the study, to document search strategy details, and to list raw data from the included studies.

8 Systematic Mapping Studies

Systematic Mapping Studies (also known as Scoping Studies) are designed to provide a wide overview of a research area, to establish if research evidence exists on a topic and provide an indication of the quantity of the evidence. The results of a mapping study can identify areas suitable for conducting Systematic Literature Reviews and also areas where a primary study is more appropriate. Mapping Studies may be requested by an external body before they commission a systematic review to allow more cost effective targeting of their resources. They are also useful to PhD students who are required to prepare an overview of the topic area in which they will be working. As an example of a mapping study see Bailey et al.'s mapping study which aimed at investigating the extent to which software design methods are supported by empirical evidence [3].

The main differences between a mapping study and systematic review are:

- Mapping studies generally have broader research questions driving them and often ask multiple research questions.
- The search terms for mapping studies will be less highly focussed than for systematic reviews and are likely to return a very large number of studies, for a mapping study however this is less of a problem than with large numbers of results during the search phase of the systematic review as the aim here is for broad coverage rather than narrow focus.
- The data extraction process for mapping studies is also much broader than the data extraction process for systematic reviews and can more accurately be termed a classification or categorisation stage. The purpose of this stage is to classify papers with sufficient detail to answer the broad research questions and identify papers for later reviews without being a time consuming task.
- The analysis stage of a mapping study is about summarising the data to answer the research questions posed. It is unlikely to include in depth analysis techniques such as meta-analysis and narrative synthesis, but totals and summaries. Graphical representations of study distributions by classification type may be an effective reporting mechanism.
- Dissemination of the results of a mapping study may be more limited than for a systematic review; limited to commissioning bodies and academic publications, with the aim of influencing the future direction of primary research.

9 Final remarks

This report has presented a set of guidelines for planning, conducting, and reporting a systematic review. The previous versions of these guidelines were based on guidelines used in medical research. However, it is important to recognise that software engineering research is not the same as medical research. We do not undertake randomised clinical trials, nor can we use blinding as a means to reduce distortions due to experimenter and subject expectations. For this reason, this version of the guidelines has incorporated information from text books authored by researchers from the social sciences.

These guidelines are intended to assist PhD students as well as larger research groups. However, many of the steps in a systematic review assume that it will be undertaken

by a large group of researchers. In the case of a single researcher (such as a PhD student), we suggest the most important steps to undertake are:

- Developing a protocol.
- Defining the research question.
- Specifying what will be done to address the problem of a single researcher applying inclusion/exclusion criteria and undertaking all the data extraction.
- Defining the search strategy.
- Defining the data to be extracted from each primary study including quality data.
- Maintaining lists of included and excluded studies.
- Using the data synthesis guidelines.
- Using the reporting guidelines

In our experience this "light" version of a systematic review is manageable for PhD students. Furthermore, research students often find the well-defined nature of a systematic review helpful both for initial scoping exercises and for more detailed studies that are necessary to position their specific research questions.

10 References

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

[2] 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.

[3] Bailey, J., Budgen, D., Turner, M., Kitchenham, B., Brereton, P. and Linkman, S. Evidence relating to Object-Oriented software design: A survey. ESEM07.

[4] Berlin, J.A., Miles, C.G., Crigliano, M.D. Does blinding of readers affect the results of meta-analysis? Online J. Curr. Clin. Trials, 1997: Doc No 205.

[5] Brereton, Pearl , Kitchenham, Barbara A., Budgen, David, Turner, Mark and Khalil, Mohamed. Lessons from applying the systematic literature review process within the software engineering domain. JSS 80, 2007, pp 571-583.

[6] Budgen, David, Stuart Charters, Mark Turner, Pearl Brereton, Barbara Kitchenham and Stephen Linkman Investigating the Applicability of the Evidence-Based Paradigm to Software Engineering, Proceedings of WISER Workshop, ICSE 2006, 7-13, May 2006, ACM Press.

[7] Cochrane Collaboration. Cochrane Reviewers' Handbook. Version 4.2.1. December 2003

[8] Cochrane Collaboration. The Cochrane Reviewers' Handbook Glossary, Version 4.1.5, December 2003.

[9] Cohen, J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Pychol Bull (70) 1968, pp. 213-220.

[10] Crombie, I.K. The Pocket Guide to Appraisal, BMJ Books, 1996.

[11] Fink, A. Conducting Research Literature Reviews. From the Internet to Paper, Sage Publication, Inc., 2005.

[12] Greenhalgh, Trisha. How to read a paper: The Basics of Evidence-Based Medicine. BMJ Books, 2000.

[13] Hart, Chris. Doing a Literature Review. Releasing the Social Science Research Imagination. Sage Publications Ltd., 1998.

[14] Jasperson, Jon (Sean), Butler, Brian S., Carte, Traci, A., Croes, Henry J.P., Saunders, Carol, S., and Zhemg, Weijun. Review: Power and Information Technology Research: A Metatriangulation Review. MIS Quarterly, 26(4): 397-459, December 2002.
[15] Jørgensen, M., and Shepperd, M. A Systematic Review of Software Development Cost Estimation Studies IEEE Transactions on SE, 33(1), 2006, pp33-53.
[16] Jørgensen, M.A review of studies of expert estimation of software development effort, Journal of Systems and Software, 70, 2002, pp 37-60.
[17] Jørgensen, M. Estimation of Software Development Work Effort: Evidence on Expert Judgment and Formal Models, International Journal of Forecasting, 2007.
[18] Jørgensen, M. Evaluation of guidelines for performing systematic literature reviews in software engineering, version 2.2, 2007
[19] 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.
[20] Khan, Khalid, S., Kunz, Regina, Kleijnen, Jos and Antes, Gerd. Systematic Reviews to Support Evidence-based Medicine, The Royal Society of Medicine Press Ltd., 2003.
[21] Kitchenham, B., Mendes, E., Travassos, G.H. (2007) A Systematic Review of Cross- vs. Within-Company Cost Estimation Studies, IEEE Trans on SE, 33 (5), pp 316-329.
[22] Lawlor, Debbie A., George Davey Smith, K Richard Bruckdorfer, Devi Kundu, Shah. Ebrahim Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? The Lancet, vol363, Issue 9422, 22 May, 2004.
[23] Noblit, G.W. and Hare, R.D. Meta-Ethnography: Synthesizing Qualitative Studies. Sage Publications, 1988.
[24] Pai, Madhukar., McCulloch, Michael., Gorman, Jennifer D., Pai, Nitika, Enanoria, Wayne, Kennedy, Gail, Tharyan, Prathap, and Colford, John, M. Jr. Systematic reviews and meta-analyses: An illustrated, step-by-step guide, The National Medical Journal of India, 17(2), 2004, pp 84-95.
[25] Petticrew, Mark and Helen Roberts. Systematic Reviews in the Social Sciences: A Practical Guide, Blackwell Publishing, 2005, ISBN 1405121106
[26] Shadish, W.R., Cook, Thomas, D. and Campbell, Donald, T. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, 2002.
[27] Staples, M. and Niazi, M. Experiences using systematic review guidelines. Article available online, JSS.
[28] Sutcliffe, T.J., Harden, K., Oakley, A., Oliver, A., Rees,S., Brunton, R. and Kavanagh, G. Children and Healthy Eating: A systematic review of barriers and facilitators, London, EPPI-Centre, Social Science Research Unit, Institute of Education, University og London, October 2003.
[29] Turner, M., Kitchenham, B., Budgen, D., Charters, S. and Brereton, P. A Systematic Literature Review of the technology Acceptance Model and its Predictive Capabilities, Keele University and University of Durham Joint Technical Report, 2007.

Appendix 1 Steps in a systematic review

Guidelines for systematic review in the medical domain have different view of the process steps needed in a systematic review. The Systematic Reviews Group (UC Berkeley) present a very detailed process model [24], other sources present a coarser process. These process steps are summarised in Table 9, which also attempts to collate the different processes.

Table 9 Systematic review process proposed in different sources

Systematic Reviews Group [24]	Australian National Health and Medical Research Council [1]	Cochrane Reviewers Handbook [7]	CRD Guidance [19]	Petticrew and Roberts [25]	Fink [11]
			Identification of the need for a review. Preparation of a proposal for a systematic review		
Define the question & develop draft protocol Identify a few relevant studies and do a pilot study; specify inclusion/exclusion criteria, test forms and refine protocol.		Developing a protocol	Development of a review protocol		
	Question Formulation	Formulating the problem		Refine questions and define boundaries	Select Research Questions
Identify appropriate databases/sources.	Finding Studies	Locating and selecting studies for reviews	Identification of research	Define Inclusion/Exclusion criteria	Select Bibliographic Databases and Web Sites. Choose Search Terms
Run searches on all relevant data bases and sources. Save all citations (titles/abstracts) in a reference manager. Document search strategy			Selection of studies	Find the primary studies	Find the studies

Researchers (at least 2) screen titles & abstracts. Researchers meet & resolve differences. Get full texts of all articles. Researchers do second screen. Articles remaining after second screen is the final set for inclusion					Apply Practical Screening criteria
Researchers extract data including quality data. Researchers meet to resolve disagreements on data Compute inter-rater reliability. Enter data into database management software	Appraisal and selection of studies	Assessment of study quality	Study quality assessment	Assess study quality	Apply methodological Quality Screen
		Collecting data	Data extraction & monitoring progress		Train Reviewers Pilot the Reviewing Process Do the Review
Import data and analyse using meta-analysis software. Pool data if appropriate. Look for heterogeneity.	Summary and synthesis of relevant studies	Analysing & presenting results	Data synthesis	Synthesize the evidence. Explore heterogeneity and publication bias	Synthesize the results Produce a descriptive review or perform meta-analysis
Interpret & present data. Discuss generalizability of conclusions and limitations of the review. Make recommendations for practice or policy, & research.	Determining the applicability of results. Reviewing and appraising the economics literature.	Interpreting the results	The report and recommendations. Getting evidence into practice.	Disseminate the results	

Appendix 2 Software Engineering Systematic Literature Reviews

Software engineering SLRs published between 2004 and June 2007 that scored 2 or more on University of York, CRD DARE scale as assessed by staff working on the Keele University and Durham University EBSE project.

Author	Fecha	Title	Reference Details	Topic type	Topic area	Quality Score
Barcelos, R.F., and Travassos, G.H.	2006	Evaluation approaches for Software Architectural Documents: A systematic Review	Ibero-American Workshop on Requirements Engineering and Software Environments (IDEAS). La Plata, Argentina.	Technology evaluation	Software Architecture Evaluation Methods	2.5
Dyba, T; Kampenes, V.B. and Sjoberg, D.I.K..	2006	A systematic review of statistical power in software engineering experiments	Information and Software Technology, 48(8), pp 745-755.	Research trends	Power in SE Experiments	2.5
Glass, R.L., Ramesh, V., and Vessey, I	2004	An Analysis of Research in Computing Disciplines	CACM, Vol. 47, No. 6, pp89-94.	Research Trends	Comparative trends in CS, IS and SE	2
Grimstad, S., Jorgensen, M. and Molokken-Ostvold, K	2006	Software effort estimation terminology: The tower of Babel	Information and Software Technology, 48 (4), pp 302-310	Technology	Cost Estimation	3
Hannay, J E., Sjøberg, D.I.K and Dybå. T	2007	A Systematic Review of Theory Use in Software Engineering Experimentos	IEEE Trans on SE, 33 (2), pp 87-107.	Research trends	Theory in SE Experimentos	2.5
Jørgensen, M	2004	A review of studies on expert estimation of software development effort,	Journal of Systems and Software, 70 (1-2), pp37-60.	Technology	Cost Estimation	3
Jørgensen, M., and Shepperd, M.	2007	A Systematic Review of Software Development Cost Estimation Studies	IEEE Transactions on SE, 33(1), pp33-53.	Research trends	Cost Estimation	3
Kampenes, V.B., Dybå, T., Hannay, J.E. and Sjøberg, D.I.K. (	2007	A systematic review of effect size in software engineering experiments.	Information and Software Technology, In press.	Research trends	Effect size in SE experiments	2.5
Mair, C. and Shepperd, M.	2005	The consistency of empirical comparisons of regression and analogy-based software project cost prediction	International Symposium on Empirical Software Engineering	Technology evaluation	Cost Estimation	2
Mendes, E.	2005	A systematic review of Web engineering research.	International Symposium on Empirical Software Engineering	Research Trends	Web Research	2

Moløkken-Østvold, K.J., Jørgensen, M. Tanilkan, S.S., Gallis,H., Lien, A.C. and Hove, S.E.	2004	Survey on Software Estimation in the Norwegian Industry	Proceedings Software Metrics Symposium.	Technology evaluation	Cost Estimation	2
Petersson, H., Thelin, T, Runeson, P, and Wohlin, C.	2004	Capture-recapture in software inspections after 10 years research – theory, evaluation and application	Journal of Systems and Software, 72, 2004, pp 249-264	Technology evaluation	Capture-recapture in Inspections	2.5
Runeson, P., Andersson, C., Thelin, T., Andrews, A. and Berling, T.	2006	What do we know about Defect Detection Methods?	IEEE Software, 23(3) 2006, pp 82-86.	Technology evaluation	Testing methods	2
Sjoeberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.K. and Rekdal, A.C.	2005	A survey of controlled experiments in software engineering	IEEE Transactions on SE, 31 (9), 2005, pp733-753.	Research trends	SE experiments	2
Zannier, C, Melnick, G. and Maurer, F.	2006	On the Success of Empirical Studies in the International Conference on Software Engineering	ICSE06, pp 341-350	Research Trends	Empirical studies in ICSE	3.5

Appendix 3 Protocol for a Tertiary study of Systematic Literature Reviews and Evidence-based Guidelines in IT and Software Engineering

Barbara Kitchenham, Pearl Brereton, David Budgen, Mark Turner, John Bailey and Stephen Linkman

Antecedentes

At ICSE04, Kitchenham et al. (2004) Suggested software engineering researchers should adopt "Evidence-based Software Engineering" (EBSE). EBSE aims to apply an evidence-based approach to software engineering research and practice. The ICSE paper was followed-up by a paper at Metrics05 (Jørgensen et al., 2005) and an article in IEEE Software (Dybå et al., 2005).

Following these papers, staff at the Keele University School of Computing and Mathematics proposed a research project to investigate the feasibility of EBSE. This proposal was funded by the UK Economics and Physical Science Research Council (EPSRC). The proposal was amended to include the Department of Computer Science, University of Durham when Professor David Budgen moved to Durham. The EPSRC have now funded a joint Keele and Durham follow-on project (EPIC).

The purpose of the study described in this protocol is to review the current status of EBSE since 2004 using a tertiary study to review articles related to EBSE in particular articles describing Systematic Literature reviews (SLRs)

Evidence-based research and practice was developed initially in medicine because research indicated that expert opinion based medical advice was not as reliable as advice based on scientific evidence. It is now being adopted in many domains e.g. Criminology, Social policy, Economics, Nursing etc. Based on Evidence-based medicine, the goal of Evidence-based Software Engineering is:

> "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." (Dybå et al., 2005)

In this context evidence is defined as a synthesis of best quality scientific studies on a specific topic or research question. The main method of synthesis is a Systematic Literature Review (SLR). In contrast to an ad hoc literature review, an SLR is a methodologically rigorous review of research results.

Research Questions

The research questions to be addressed by this study are:

- How much EBSE activity has there been since 2004?
- What research topics are being addressed?
- Who is leading EBSE research?

- What are the limitations of current research?

Search Process

The search process is a manual search of specific conference proceedings and journal papers since 2004. The nominated journals and conferences are shown in the following Table.

Sources to be Searched

Fuente	Responsible
Information and Software Technology (IST)	Kitchenham
Journal of Systems and Software	Kitchenham
IEEE Transactions on Software Engineering	Kitchenham
IEEE Software	Kitchenham
Communications of the ACM (CACM)	Brereton
ACM Surveys	Brereton
Transactions on Software Engineering Methods (TOSEM)	Brereton
Software Practice and Experience	Budgen & Kitchenham
Empirical Software Engineering Journal (ESEM)	Budgen
IEE Proceedings Software (now IET Software)	Kitchenham
Proceedings International Conference on Software Engineering (ICSE 04, 05, 06, 07)	Linkman & Kitchenham & Brereton
Proceedings International Seminar of Software Metrics (Metrics04, Metrics05)	Kitchenham & Brereton
Proceedings International Seminar on Empirical Software Engineering (ISESE 04, 05, 06)	Kitchenham & Brereton

Specific researchers will also be contacted directly:
Dr Magne Jørgensen
Professor Guilherme Travassos.

Inclusion criteria

Articles on the following topics, published between Jan 1st 2004 and June 30th 2007, will be included

- Systematic Literature Reviews (SLRs) i.e. Literature surveys with defined research questions, search process, data extraction and data presentation
- Meta-analyses (MA)

Exclusion Criteria

The following types of papers will be excluded

- Informal literature surveys (no defined research questions, no search process, no defined data extraction or data analysis process).
- Papers discussing process of EBSE.
- Papers not subject to peer-review.

When an SLR has been published in more than one journal/conference, the most complete version of the survey will be used.

Primary study selection process

The results will be tabulated as follows:

- Number of papers per year per source
- Number of candidate papers per year per source
- Number of selected papers per year per source.

The relevant candidate and selected studies will be selected by a single researcher. The rejected studies will be checked by another researcher. We will maintain a list candidate papers that were rejected with reasons for the rejection.

Quality Assessment

Each SLR will be evaluated using the York University, Centre for Reviews and Dissemination (CDR) Database of Abstracts of Reviews of Effects (DARE) criteria (http://www.york.ac.uk/inst/crd/crddatabase.htm#DARE). The criteria are based on four questions:

- ¿Se describen y son adecuados los criterios de inclusión y exclusión del examen?
- ¿Es probable que la búsqueda bibliográfica haya abarcado todos los estudios pertinentes?
- ¿Los revisores evaluaron la calidad/validez de los estudios incluidos?
- ¿Se describieron adecuadamente los datos/estudios básicos?

The questions are scored as follows:

- Question 1: Y (yes), the inclusion criteria are explicitly defined in the paper, P (Partly), the inclusion criteria are implicit; N (no), the inclusion criteria are not defined and cannot be readily inferred.
- Question 2: Y, the authors have either searched 4 or more digital libraries and included additional search strategies or identified and referenced all journals addressing the topic of interest; P, the authors have searched 3 or 4 digital libraries with no extra search strategies, or searched a defined but restricted set of journals and conference proceedings; N, the authors have search up to 2 digital libraries or an extremely restricted set of journals.
- Question 3: Y, the authors have explicitly defined quality criteria and extracted them from each primary study; P, the research question involves quality issues that are addressed by the study; N no explicit quality assessment of individual papers has been attempted.

- Question 4: Y Information is presented about each paper; P only summary information is presented about individual papers; N the results of the individual studies are not specified.

The scoring procedure is Y=1, P=0.5 and N or Unknown=0.

The data will be extracted by one researcher and checked by another.

Data Collection

The data extracted from each paper will be:

- The source (i.e. the conference or journal).
- The year when the paper was published. Note if the paper was published in several difference sources both dates will be recorded and the first date will be used in any analysis. This is necessary in order to track the EBSE activity over time.
- Classification of paper
    - Type (Systematic Literature Review SLR, Meta-Analysis MA).
    - Scope (Research trends or specific research question).
- Main software engineering topic area.
- The author(s) and affiliation (organisation and country).
- Research question/issue.
- Whether the study referenced an EBSE paper or the SLR Guidelines (Kitchenham, 2004).
- Whether the study resulted in evidence-based practitioner guidelines.
- The number of primary studies used in the SLR/MA
- Summary of paper.
- Quality score for the study.

The data will be extracted by one researcher and checked by another.

Data Analysis

The data will be tabulated (ordered alphabetically by the first author name) to show the basic information about each study. The number of studies in each major category will be counted.

The tables will be reviewed to answer the research questions and identify any interesting trends or limitations in current EBSE-related research as follows:

- Question 1 How much EBSE activity has there been since 2004? This will be addressed by simple counts of the number of EBSE related papers per year.
- Question 2 What research topics are being addressed? This will be addressed by counting the number of papers in each topic area. We will identify whether any specific topic areas that have a relatively large number of SLRs.
- Question 3 Who is leading EBSE research? We will investigate whether any specific organisation of researches have undertaken a relatively large number of SLRs.
- Question 4 What are the limitations of current research? We will review the range of SE topics, the scope of SLRs and the quality of SLRs to determine

whether there are any observable limitations. We will also investigate whether the quality of studies is increasing over time by plotting the quality score against the first publication date, and whether the quality of studies has been influenced by the SLR guidelines (by comparing the average quality score of SLRs that referenced the guidelines with the average score of SLRs that did not reference the guidelines).

Dissemination

The results of the study should be of interest to the software engineering community as well as researchers interested in EBSE. For that reason we plan to report the results on a Web page. We will also document the full result of the study in a joint Keele University and University of Durham technical report. A short version of the study will be submitted to IEEE Software.

Referencias

1. Barbara Kitchenham, Tore Dybå and Magne Jørgensen. (2004) Evidence-based Software Engineering. Proceedings of the 26th International Conference on Software Engineering, (ICSE ’04), IEEE Computer Society, Washington DC, USA, pp 273 – 281 (ISBN 0-7695-2163-0).
2. Kitchenham, B. Procedures for Performing Systematic Reviews. Joint Technical Report, Keele University TR/SE-0401 and NICTA 0400011T.1, July 2004.
3. Tore Dybå, Barbara Kitchenham, and Magne Jørgensen. Evidence-based Software Engineering for Practitioners, IEEE Software, Volume 22 (1) January, 2005, pp58-65.
4. Magne Jørgensen, Tore Dybå, and Barbara Kitchenham. Teaching Evidence-Based Software Engineering to University Students, 11th IEEE International Software Metrics Symposium (METRICS'05), 2005, p. 24.