NECA 연구방법 시리즈 7

한국적 상황을 고려한 비교효과연구 방법

A Summary of Methods for Comparative Effectiveness Research

Version 1.0

NECA 한국보건의료연구원
National Evidence-based Healthcare Collaborating Agency

NECA 연구방법 시리즈 7

한국적 상황을 고려한 비교효과연구 방법

A Summary of Methods
for Comparative Effectiveness Research

Version 1.0

NECA 연구방법 시리즈 7
한국적 상황을 고려한 비교효과연구 방법

인쇄: 2013년 2월 28일 초판 1쇄 발행
발행인: 이선희
발행처: 한국보건의료연구원
서울시 종로구 원남동 율곡로 174 창경빌딩
Tel. 02-2174-2700
Homepage: www.neca.re.kr

ISBN 978-89-6834-002-4
ISBN 978-89-6834-003-1 (세트)

First edition.

Published by NECA
February 2013

Comments and suggestions on a summary of methods for comparative effectiveness research can be made at www.neca.re.kr

집필진

본 책자는 다음과 같이 인용될 수 있습니다.

안정훈, 김윤희, 이향열, 장보형, 장은진, 현민경, 김윤정, 안지혜, 조송희
한국적 상황을 고려한 비교효과연구 방법. 한국보건의료연구원. 2013.

Suggested citation:

J Ahn, Y Kim, HY Lee, BH Jang, EJ Jang, MK Hyun, YJ Kim, JH An, and SH Cho.
A summary of methods for Comparative Effectiveness Research

안정훈 연구위원
한국보건의료연구원 보건서비스분석실

김윤희 책임연구원
한국보건의료연구원 보건서비스분석실

이향열 책임연구원
한국보건의료연구원 보건서비스분석실

장은진 책임연구원
한국보건의료연구원 보건서비스분석실

장보형 책임연구원
한국보건의료연구원 보건서비스분석실

현민경 책임연구원
한국보건의료연구원 보건서비스분석실

김윤정 연구사
한국보건의료연구원 보건서비스분석실

안지혜 연구사
한국보건의료연구원 보건서비스분석실

조송희 연구사
한국보건의료연구원 보건서비스분석실

발간사

국민에게 양질의 의료를 제공하고 의료자원을 적절하게 활용하는 것은 대다수 국가들이 주요 의료정책 과제로 추구하고 있는 현안이며, 우리나라에서도 지속적인 정책과제가 되어 왔습니다. 특히 최근엔 국제적으로 한정된 의료자원을 효과적으로 투입하여 국민 건강을 향상시키기 위한 전략으로서 과학적 · 합리적인 근거를 기반으로 한 정책의사결정에 대해 관심이 높아지고 있습니다. 국내에서도 이에 대한 필요성과 공감대가 높아지고는 있으나 근거중심 보건의학을 위한 체계적이고 통일된 방법은 수립되지 못하고 있고 연구수준도 걸음마 단계에 놓여있는 실정입니다.

한국보건의료연구원에서는 개원 이래 지금까지 보건의료분야의 현안들에 대한 의료기술평가연구들을 진행하여 연구 결과를 실제 보건의료 정책 의사결정에 활발하게 적용할 수 있도록 노력을 지속해 왔습니다. 또한 국내에서 우리 현실에 적합한 접근 방안을 고민할 필요가 있다는 판단 하에 의료기술평가연구를 진행하는데 있어 여러 필수적인 연구방법론들의 표준화에 노력하는 한편, 활용을 촉진시키기 위한 기관 차원의 다양한 노력을 지속해 왔습니다.

그리고 그간 축적된 연구경험과 고민들을 함께 공유하고 논의하기 위한 노력의 일환으로 한국보건의료연구원에서는 「체계적 문헌고찰 매뉴얼」, 「임상진료지침 개발 매뉴얼」, 「신의료기술평가 체계적 문헌고찰 지침」을 발간한 바 있습니다. 이를 필두로 국내 연구자들에게 더 분명하고 일반화 가능한 연구 결과를 도출하기 위해 필

요하다고 여겨지는 연구방법론들을 정리하여 5권의 NECA 연구방법 시리즈를 발간하게 되었습니다. 이 연구방법 시리즈는 베이지안 메타분석법, 보건의료분야에서 비용 산출방법, 국내 보건의료 이차자료원 활용, 측정된 교란요인을 고려한 성과분석 방법, 한국적 상황을 고려한 비교효과연구 방법으로 구성되어 있습니다.

NECA 연구방법론 시리즈가 연구결과를 일반화하고 비뚤림없이 적용가능한 정보를 제공하는데 밑바탕이 되어 과학적 · 합리적인 근거를 제시하는데 도움이 되기를 바랍니다. 아울러 이번 연구방법론 시리즈 출간을 계기로 국내 연구여건에 부합된 방법론 정립과 활성화를 위한 많은 논의와 시도가 이루어지기를 기대하며, 이로 인해 보건의료정책의 합리성을 높일 수 있는 근거창출 연구기반이 더욱 공고해지기를 바랍니다.

2013년 2월

한국보건의료연구원장 이 선 희

머리말

한국은 인구의 고령화와, 만성질환, 암질환, 사고 등의 증가로 인해 의료비 지출이 급격히 증가되어 효율적인 자원의 이용에 대한 사회적 요구가 증가하고 있다. 또한, 의료기술의 전문화로 인해 진료의 종류가 다양해지면서, 의료에 종사하는 의료진뿐만 아니라, 의료를 제공받기를 원하는 소비자들이 무엇이 최선의 진료인지, 내가 받는 진료가 과연 최선인지에 대한 관심도 증가하고 있다. 이러한 상황 속에서 비교효과연구(Comparative Effectiveness Research, CER)는 보건의료의 실무를 발전시키려는 사회의 거대한 메커니즘 속에 하나의 중요한 연구방법으로서 주목을 받고 있다.

비교효과연구는 보건의료전달체계를 향상시키기 위해 어떠한 임상적 상황(clinical condition)을 예방하고, 진단하고, 치료하고, 모니터링하기 위한 중재들(interventions)과 대안적 방법들(alternative methods)의 이득(benefits)과 위해(harms)를 비교하기 위해 근거를 생성하고 합성하는 것을 말한다. 비교효과연구는 환자, 의료제공자, 정책결정자, 보험지불자와 같은 의사결정자들에게 어떤 중재들이 특정 유형의 환자들에서 가장 효과적인지에 대한 정보를 제공함으로써 환자의 치료성과를 향상시키는 것을 목표로 한다. 미국에서 처음 각광을 받게 된 비교효과연구가 최근 한국에서도 관심의 대상이 되는 이유는 더 좋은 정보로 인하여 더 나은 치료결과를 가져오거나 의료자원을 더 효과적으로 활용할 것이라는 믿음 때문일 것이다.

본 책자는 6개의 장으로 구성되어 있다. 1장은 비교효과 임상시험, 2장은 비교효과 관찰연구, 3장은 체계적 문헌고찰, 4장은 메타분석, 5장은 비용-효과분석, 6장은 특정분야의 비교효과연구에 대한 방법론에 대해 기술하였다. 비교효과연구를 수행하기 위한 이론적 배경을 개요에 담았으며, 실제로 연구질문을 가지고 어떠한 접근방법을 활용해야 하는지에 대한 판단을 할 수 있도록 다양한 비교효과연구의 세부방법론에 대한 표준적 접근법을 제시하였다. 실제 상황에서 비교효과연구의 결과를 건설적으로 의사결정에 활용하기 위해서는 환자나 의료진들은 진료의 계획을 수정할 때 어느 정도의 불확실성을 감당해야 한다. 하지만 표준적인 연구방법론으로 명쾌하게 연구 질문에 대한 해답을 제공할 수 있다면 명확한 근거를 바탕으로 합리적인 의사결정을 내릴 수 있을 것이다.

본 책자는 한국적 상황을 고려한 비교효과연구의 초석이라 할 수 있다. 본 책자를 통해 비교효과연구가 한국에서 활발하게 수행되고, 그 연구결과들을 통해 한국의 보건의료의 질 향상에 활용할 수 있게 되길 기대해 본다.

2013년 2월

일러두기

1. 「한국적 상황을 고려한 비교효과연구 방법」은 비교효과연구 수행시 필요한 연구 방법을 한국적 상황에 맞게 정리한 한국보건의료연구원의 공식 책자입니다.

2. 본 책자는 한국보건의료연구원 홈페이지에서 다운로드가 가능합니다.
(한국보건의료연구원 홈페이지 http://www.neca.re.kr)

3. 본 책자는 주기적으로 개정될 예정이며 본 책자와 관련된 문의 및 건의사항은 한국보건의료연구원(neca@neca.re.kr)으로 해주시기 바랍니다.

CONTENTS

PART 1

비교효과 임상시험

일반적인 무작위 배정 임상시험(Randomized Controlled Trials, RCT)은 치료군과 대조군 사이에서 효능(efficacy)을 보기위한 연구이다. 미국 AHRQ(Agency for Healthcare Research and Quality)에서 발행한 「비교효과연구 방법론 가이드」에 따르면 무작위 배정 임상시험은 다른 방법에 비해 명확한 결과를 얻을 수 있는 장점이 있지만 비교효과연구에서는 RCT에만 의존한 방법론은 문제가 있다고 지적하고 있다. 비교효과연구에서의 RCT방법은 현실을 반영하기에 연구기간이 너무 짧고, 최소 1천만 달러이상의 비용이 들며, 효과보다는 효능을 보는 연구들이 대부분이다. 또한, RCT연구는 대상자수가 적으며, 중간결과보다는 주요 건강관련 결과만을 보고하고 있고, 부작용은 잘 보고가 되지 않거나 누락되어, 출판으로 인한 연구결과의 편향, 선택적 결과보고 등의 바이어스를 초래할 수 있다. 이러한 이유로 RCT는 비교효과연구의 방법론으로 비효율적일 수 있다.

이러한 단점에도 불구하고 비교효과연구에서 RCT는 직접비교 임상시험일 경우 치료제들을 1:1로 직접 비교할 수 있는 근거를 제공할 수 있고, 비교효과연구에서 고려해야 할 위약, 대조군 등에 대한 정보도 제공하기 때문에 일정 부분 활용할 가치가 있다(AHRQ, 2007).

비교효과연구에서의 RCT방법을 활용하기 위해서는 단순히 효능을 본 임상시험이 아니라 알려진 치료보다 더 도움이 되는 상위의 치료를 선택하고 결정할 수 있도록 설계되어야 한다. 치료의 선택은 다른 약물, 다른 수술 기술 등과 같이 동일 치료법 내에서 다른 약물끼리, 혹은 기술끼리 비교할 수도 있고, 약물과 수술, 수술과 의료기기, 세심한 관찰과 즉각적인 중재, 행동요법과 약물요법 등과 같이 각기 다른 치료법의 그룹 간 서로 비교할 수도 있다(Peduzzi 등, 2010).

비교효과연구를 위해서 RCT연구는 연구설계와 분석 측면에서 관점의 전환이 필요하다. 비교효과연구는 환자, 소비자, 임상전문가들에게 유용한 정보뿐만이 아니라 보험자, 정책결정자에게 의사결정을 위한 정보를 제공하는 것을 목적으로 하고 있기 때문에, 적응임상시험(adaptive clinical trial) 설계와 베이지안 분석방법을 활용하거나 실제 임상환경에서의 임상시험인 실용임상시험(pragmatic clinical trial 또는 practical clinical trial)을 활용하는 등의 전환을 해야 한다. 적응임상시험 방법은 기존의 RCT에 비해 대상자수, 시간, 비용을 줄일 수 있고, 하위그룹을 정의하고 분석이 가능하며, 1:1 뿐만 아니라 다양한 치료방법을 비교할 수 있다. 또한, 적응임상방법은 기존의 높은 질의 외부 근거(체계적 문헌고찰, 명확하게 증명된 관찰연구 등)를 포함하여 의사결정에 반영할 수 있는 장점이 있다. 실용임상연구는 실제 임상환경에서 수행한 치료제 및 치료전략(예. 병합요법 약물)의 효과(effectiveness)를 비교하는 연구로서 비교효과연구에 적합한 연구설계라 할 수 있다 (Bryan

등, 2009).

일반적인 RCT는 이상적인 환경에서 위약을 주로 대조군으로 하고 탐색적인 결과를 포함하고 있다. 동일한 선정기준과 대학 또는 특정 의료기관을 포함하고 즉시 도출가능한 결과 또는 대리지표를 결과로 사용한다. 그러나 비교효과연구는 실제 임상환경과 일반적인 환자를 대상으로 보다 현실적인 결과를 도출하고 기존의 표준치료법과 새로운 치료법을 비교해 보다 효과적인 치료법을 찾을 수 있는 장점이 있다.

이처럼 비교효과연구에서 임상시험을 활용할 경우 몇 가지 고려해야 할 사항은 1) 위험요인 관리, 2) 임상적 균형, 3) 환자선호도, 4) 최신 기술, 5) 비교대안으로 일반치료 등이다. 위험요인 관리는 임상시험을 설계할 때 단일 질환만을 대상으로 하는 경우가 많지만 노인 등 피험자들이 하나 이상의 동반질환을 가지고 있는 경우가 많고 동반질환이 치료효과에 영향을 주지 않도록 조정해야 한다. 광범위 임상연구를 시행할 경우 지역마다 선호되는 치료법이 있기 때문에 이들 간의 임상적 균형을 유지하는 것이 중요하다. 이는 치료 근거에 대한 의견을 나누고 치료 패턴 등을 정리, 확립함으로써 일부 해결할 수 있다. 환자선호도도 간과되어서는 안 되는 부분이다. 층화 무작위 방법을 통해서 환자의 치료선호도가 균형을 이룰 수 있도록 조정하거나 환자중심의 접근법을 활용해 개별적인 치료선호도 등을 고려해 설계하고 이들 선호가 연구실행 및 타당성에 어느 정도 영향을 미치는 지를 파악하여 설계해야 한다.

최신의 기술을 항상 연구에 적용되도록 설계하는 것이 항상 옳은 것은 아니다. 이상적으로는 잠재적 발전 기술은 연구설계 단계에서부터 파악되어야 하지만 항상 가능한 것은 아니다. 임상시험에 새로운 기술을 수용하기 위해서는 관련 임상전문가들을 대상으로 연구초기부터 확실하게 교육하고, 반복적으로 교육을 해야 하며, 연구설계시 무작위배정 원칙하에 연구가 진행되도록 해야 한다. 또한 임상시험에서 비교대안으로 일반치료군을 다룰 때는 환자의 이익과 위험을 고려하고, 과학적 검증과 임상현장에서의 타당성과 현실가능성을 바탕으로 설계해야 한다(Peduzzi 등, 2010).

비교효과 임상시험의 초기 예로 안정성 협심증(stable angina)에서 관상동맥 우회술 관련 무작위 임상시험을 들 수 있다. 이 임상시험은 관상동맥 우회술이 개발된 후 바로 시행되었다. 임상시험의 목적은 안정성 협심증 환자에서 새로운 기술과 기존의 약물 치료방법을 비교하는 것이었다. 임상시험의 결과 대부분의 환자에서 약물치료가 안전할 수 있고 수술을 해야 할 정도로 진행되거나 증상이 있는 고위험 환자에서만 관상동맥 우회술이 유용한 것으로 나타났다. 이 임상시험 결과는 임

상현장에 반영되어 관상동맥 우회술 감소를 이끌었다(Detre 등, 1984; Peduzzi 등, 2010).

COURAGE 임상시험은 만성 허혈성 심장질환의 최선의 치료전략을 결정하기 위한 연구였다. 약물치료와 생활습관개선 등의 치료를 받은 환자와 경피적 관상동맥 중재술(Percutaneous Coronary Intervention, PCI)과 함께 약물치료를 받은 환자를 비교하였다. 미국과 캐나다 50개 센터에서 2,287명의 환자가 등록되었다. COURAGE 임상시험에서 PCI와 약물요법을 병행했을 때 초기 치료전략으로 효과적이라는 것을 증명했지만 사망과 심근경색 위험을 감소시키지는 않았다. 이러한 결과를 바탕으로 안정적인 만성 허혈성 심장질환 환자에서는 먼저 약물요법을 하고 필요시 관상동맥 중재술을 받을 수 있다는 내용을 임상진료지침에 반영하였다(Boden 등, 2006; Boden 등, 2007; Peduzzi 등, 2010).

1.1. 군집 무작위 배정 시험

1.1.1. 개요

군집 무작위 배정 시험(Cluster Randomized Trials, CRT)은 중재군과 대조군을 배정하는 방식에 있어서 환자 개개인이 아니라 의료제공자, 병원, 지역사회 보건의료체계단위 등과 같이 그룹 또는 군집을 대상으로 무작위 배정하는 방법이다(Cambell and Grimshaw, 1998; Donner 1998; Puffer 등, 2005; Eldridge 등, 2008). CRT는 제어된 환경에서의 임상시험인 RCT와는 달리 실제 임상환경 하에서 치료의 효과를 평가할 때 주로 사용한다(Mazor 등, 2007; Platt 등, 2010).

RCT는 엄격한 표준화가 요구되고, 일반 임상현장에서 사용되는 의학적 치료보다 비용이 많이 들고, 적용하는데 시간이 많이 소요된다(Platt 등, 2010). 반면 CRT는 기존의 연구 인프라를 활용하거나 대규모의 보험급여 정보를 가지고 있는 경우 적용할 수 있다. 또한 환자들의 치료 정보와 치료결과에 대한 자료를 얻는데 비용을 절감하고 시간을 절약할 수 있다(Mazor 등, 2007).

구체적으로 비교효과연구에서 사용되는 군집 무작위 배정 시험은 다음과 같은 이점이 있다(Platt 등, 2010). 첫째, CRT는 실제 진료가 이루어지는 환경에서 병원, 진료단위, 보험급여 등 군집 단위에 대하여 중재들을 쉽게 적용하여 연구할 수 있다. 예를 들면, 기존의 임상 진료 행정 시스템을 사용하여 군집 무작위 배정을 하게

되면, 의료제공자들 간에 존재하는 그룹간 의사소통의 영향을 고려한 연구 설계를 할 수 있다. 둘째, CRT는 병원 전체나 중환자실, 나아가 보다 넓은 인구집단에 적용되기 때문에 연구결과의 일반화 가능성이 높다. 셋째, CRT는 연구과정이 단순하고 연구비용이 적고 현재 사용되는 의료전달체계의 일부분을 활용할 수 있다. 전자의무기록 시스템이 도입되면서 CRT는 이것을 일반적인 건강정보 수집 시 기초상태 사정, 중재 모니터링, 결과 측정의 목적으로 활용할 수 있게 되었다.

그러나 CRT의 연구설계시 일반 RCT보다 더 많은 주의가 필요하다. 먼저 같은 중재를 받은 환자들 사이에 치료결과의 연계성을 설명해야 할 필요가 있으며, 이러한 임상시험에서 윤리적인 문제들이 있을 수 있다. 또한 무작위화의 단위와 분석단위가 다를 가능성이 있고, 무작위화 이전에 불균형적인 교란요인이 잠재되어 있을 가능성이 높기 때문이다(Campbell 등, 2004; Gynn 등, 2007; Puffer 등, 2003; Hutton, 2001). 그 외에도 군집간의 연관성이 존재하기 때문에, 개별 RCT 설계와 비교해서 더 많은 환자들이 필요하다.

따라서 군집 무작위시험은 개인 수준에서 무작위가 어려운 조건 일 때, 연구설계 및 자료수집을 도와줄 수 있는 보건의료 시스템이 갖추어져 있는 경우에 특히 유용한 방법이 될 수 있다(Gynn 등, 2007).

1.1.2. 방법

CRT 연구 시 제일 먼저 던질 수 있는 질문은 “어떤 종류의 군집이나 그룹의 무작위화가 필요한가?”이다. 여기에서 군집의 종류, 즉 무작위화의 단위는 의사들, 진료단위, 건강보험급여 유형, 또는 지역 단위 등이 될 수 있다(Glynn 등, 2007; Mazor 등, 2007). 이때 이들은 하나의 CRT연구에서 한 군집 내의 개개인은 같은 군(study arm)으로 배정된다(Mazor 등, 2007).

군집 무작위화는 하나의 중재가 하나의 그룹 전체에 적용되어야 할 때(예, 지역사회 기반의 건강 증진 계획) 필요한 무작위화 방법이다. 또한 각 개인들의 질환상태가 연관성이 높은 경우에(예, 감염성 질병에 대한 집단 면역) 사용할 수 있다(Platt 등, 2010). 치료 결과를 유의하게 향상시킬 수 있는 중재들간 비교하기 위해서는 개별 환자보다는 임상적 환경(practice site) 단위를 그룹으로 무작위 배정을 하고 이때는 보통 각 그룹에 하나의 중재를 적용한다(Sox and Goodman, 2012).

관심 있는 연구주제에 명확한 해답을 제공하기 위해 필요한 표본의 크기(sample size)는 하나의 실무단위 내에서 진료 받은 환자들의 군집간 연관성(intracluster correlation)에 따라 달라진다(Glynn 등, 2007). 중재 전에 미리 해당 연구설계의

효과 범위를 측정해보면 더 현실적인 표본수를 찾을 수 있다. 환자가 중재를 얼마나 충실하게 받을지, 중재로 인해 어떠한 이득을 얻을지, 합리적으로 그 가능성을 따져보고 사전에 연구참여자 모집 시 명확한 제한범위를 둔다면 임상시험의 타당성을 높일 수 있을 것이다.

1.1.3. 사례

Platt 등(2010)은 병원들, 건강보험 종류, 의사들의 임상 현장 네트워크를 이용하여 군집 무작위 임상시험을 사용하였다. REDUCE-MRSA trial(ID#NCT00980980)연구는 군집 무작위 임상시험의 방법론을 사용하여 병원 중환자실에서 MRSA감염 환자의 감소를 목적으로 3개의 예방적 중재들을 평가하였다. 이 연구는 메티실린에 내성이 있는 Staphylococcus Aureus (MRSA) 감염을 예방하기 위해 사용되고 있는 3가지 종류의 MRSA감염 예방법 중 하나를 중환자실에 무작위 배정하였고 참여하는 병원 내에 있는 모든 성인 중환자실은 동일한 중재법을 할당받았다. REDUCE-MRSA trial은 45개 미국병원협회(Hospital Corporation of America, HCA) 병원들을 대상으로 연구를 진행하였다. HCA는 시스템 차원의 질 향상, 감염 관리 및 예방 프로그램, 중앙화된 정보 시스템을 포함한 형태의 중앙화된 행정 구조를 가지고 있어 군집 무작위 배정 시험을 하는데 유리하였다. 3개의 중재들은 그 병원들의 기존의 원내 교육, 이행도 및 수행도 감시 프로그램들을 사용한다. 각 군(study arm)은 중환자실 수간호사, 교육 간호사들, 감염관리 및 예방 프로그램 관리자들에 의해 질 관리, 감염관리 및 예방 프로그램이 수행된다. 이 연구는 HCA의 실제 임상현장에서 일어나는 3개의 중재들의 성과에 대해 현실성 있는 평가를 제공하기 위해 CRT연구방법을 적용하여 연구의 효율성을 높인 좋은 예이다.

1.2. 실용 임상시험

1.2.1. 개요

실용임상시험(Pragmatic clinical trials, PCT)은 일상 임상 현장에서 일어나는 중재들의 위험, 이득, 비용을 비교하기 위한 연구방법이다. 실제 환경에서 치료의 유효성을 비교하기 위해 설계된 전향적 무작위 배정 연구(prospective randomized studies)이며, 전통적 임상시험에 비해 다양한 연구대상자를 포함할 수 있기 때문에 연구결과를 일반화할 수 있으며, 연구대상자의 증가로 통계적 검정력(power)이 증가하므로 비교효과연구에서 중요한 방법이다(Saag 등, 2012). 실용임상시험은 특히 의사결정자들에게 정보(근거)를 제공하기 위한 목적으로 수행되기 때문에, 비교효과연구의 목적과 일치된다. 따라서 제한적인 조건 아래 하나의 관심 치료법이 효능이 있는지를 알아보기 위해 수행되는 전통적 임상시험과는 달리, PCT는 실제로 임상현장에서 사용되고 있는 중재의 효과성, 위험요인, 비용에 대한 연구 질문들의 답을 찾기 위해 수행될 수 있다(Saag 등, 2012; Sox and Goodman, 2012). PCT의 주된 특성은 활성 대조군(active comparator)을 사용하고, 포함기준(inclusion criteria)을 넓게 설정하며, 표본수가 큰 다양한 환자집단을 통해 실제 임상현장에서 치료받는 대상의 대표성을 확보하고, 쉽게 측정 가능한 임상적 결과지표들(clinical endpoints)을 사용한다는 것이다(Saag 등, 2012).

PCT는 일반적으로 연구기간이 길며 비용이 많이 드는데, 잘 알려진 PCT 사례들인 Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack(ALLHAT)과 the Action to Control Cardiovascular Risk in Diabetes (ACCORD)은 각각 1억 3천만 불, 3억 불이 소요되었다고 알려져 있다(AcademyHealth, 2009). 하지만 이러한 비용에 비해서 임상 현장에는 크게 영향을 미치지 않았다는 비판이 있다(Saag 등, 2012). 이에 반해 임상현장에 상당히 큰 변화를 준 효율적으로 진행된 PCT의 몇 가지 예들이 있다. 그중 가장 주목할 만한 예는 이탈리아에서 수행된 GISSI로 심근경색에서의 streptokinase에 대한 연구이다(Saag 등, 2012). 이 연구에서는 이탈리아 내의 대략 90%의 관상동맥질환 집중치료실(coronary care unit)에 내원한 심근경색 환자들을 등록하였는데 연구비가 $50,000이 들었다(Taylor, 2010). GISSI연구에서는 혈전용해제를 전혀 사용하지 않은 대조군의 병원 사망률이 13%, 혈전용해제를 사용한 치료군의 병원사망률은

8.8%로 나타나 급성심근경색 환자에서 혈전용해제를 사용하는 것이 효과가 있음을 밝혔다. GISSI연구결과 발표 전에는 대부분의 센터들이 혈전용해제(thrombolysis)를 사용하고 있지 않았으나, 연구결과 발표 후 거의 70%의 센터들이 정기적으로 혈전용해제를 사용하게 되었다(Saag 등, 2012).

치료법들에 대한 비교효과연구의 필요성이 증가하고 있으나 연구비는 제한되어 있으므로, 보다 효율적으로 연구를 진행할 수 있는 비용절약형 실용임상시험을 설계하는 것은 중요하다고 할 수 있다(Saag 등, 2012).

1.2.2. 방법

MacPherson(2004)에 따르면 PCT의 연구수행절차는 다음과 같다.

1.2.2.1 적절한 연구질문 도출

연구설계는 명확하게 정의된 연구질문에 부합해야 한다. 일상 진료에서 치료의 이득을 평가하고 싶다면, 일반적인 임상 세팅과 실용연구 설계를 이용해야 한다. 실용임상시험은 하나의 중재에 대한 전체 효과(overall effectiveness)에 대한 질문에 답을 주지만, 중재의 다른 요인들이 구체적으로 어떻게 영향을 끼치는지는 연구할 수 없다. 치료가 어떤 공헌을 했는지, 환자들의 기대치들, 사용된 특정 치료 등을 포함하는 진료 전체 패키지를 검증하기 위해 실용임상시험을 사용한다. 이때 일반적으로 위약(placebo)이 아닌, 다른 치료법과 하나의 진료 패키지의 효과가 어떻게 차이가 나는지를 비교하게 된다. 실용임상시험은 정책결정자들, 의료진들이나 환자들이 두 개의 중재들 중에 어떤 선택을 해야 할 지 의사결정을 돕는 근거를 제공하는 것을 목표로 한다.

1.2.2.2 연구대상자 정의

실용임상시험에서는 참여하는 연구대상자들의 포함기준을 넓게 설정함으로써 연구결과의 일반화가능성이 높아진다. 따라서 포함기준(inclusion criteria)을 넓게 설정하기 위해, 환자들이 다른 질환이 있거나 약물을 복용하더라도 포함한다. 연구자는 환자의 상태, 일반진료 행태와 진료의뢰, 기존 치료법에 대한 환자들의 태도, 보완적 치료법의 강점과 약점 등에 따라 가장 적절한 치료법이 무엇인지 확인해야 한다. 예를 들면, 기존의 치료가 상당히 만족스럽지 못한 경우, 환자들에게 기존 치료

를 선택하도록 하는 것은 바람직하지 않다. 이때는 환자들과 의사들은 또 다른 대안적 치료법이 있는지 고려해야 한다. 이러한 복잡한 문제들은 소규모 예비연구들(small pilot studies)을 통해 연구설계 시 명확하게 논의되어야 한다.

1.2.2.3 대조군 선정

연구대상자를 선정한 후 상대적 차이를 측정할 수 있는 기준집단(reference group)을 대조군으로 선정한다. 실용임상시험에서는 위약 대조군과 눈가림을 사용하게 될 경우 임상시험에 참여하는 환자들에게 의도치 않게 해로운 영향을 끼칠 가능성이 있기 때문에 이를 사용하는 것은 적절치 않다. 예를 들면, 눈가림을 한 의료진이 잘못된 치료법을 사용했거나, 위약 그룹에 속한 환자의 반응이 부족하면 치료의 영향을 제대로 알지 못하게 된다. 그렇기 때문에 연구자는 현재 사용되고 있는 진료 중재들 중에서 대조군을 선택해 임상시험의 두 군(arms)으로 설계해야 할 필요가 있다. 이는 현재 사용되는 치료법들 사이에서 환자나 의료진들이 가장 좋은 진료를 선택하는 것을 돕는 근거를 생성하는 것이 비교효과연구의 목표이기 때문이다. 연구범위를 결정하고 윤리적으로 타당성을 확보하기 위해서는 두 개의 치료군에서 성공할 확률이 비슷해야 공정하게 효과를 비교하는 것이 되며, 그 중 어떤 중재를 받더라도 환자가 손해를 안 보게 된다.

1.2.2.4 치료 프로토콜 정의하기

실용임상시험은 각기 다른 환자들에게 복잡하고 개별적인 접근을 허용하면서, 통상적인 환자진료를 하도록 진료 시 자율성을 존중해 준다. 이는 환자의 추적관찰(follow-up) 시 의료진들이 정밀하게 관찰하고 질문하면서 치료과정에 치료법의 미세한 조종이 가능하게 하는 것을 의미한다. 이 때문에 PCT는 일상적으로 이루어지는 임상 진료환경에서 설계되고 상당히 숙련된 의료진들을 필요로 한다. 따라서 정해진 치료 프로토콜에는 어느 정도의 치료법의 변형이 가능한지 명시해야 한다. 규정된 틀 내에서 넓은 유연성을 가지는 굉장히 열린 프로토콜에서부터 전문가들의 합의로 결정된 엄격한 프로토콜까지 여러가지 옵션이 있을 수 있다. 복잡한 중재들이 있다면 첫째로 치료에 필요한 지표들을 정의하는 핸드북이나 매뉴얼을 만들어야 한다. 연구의 프로토콜은 반복가능(replicated)하고, 동시에 일반화가능(generalizable)하여 일상진료에 사용할 수 있어야 한다.

1.2.2.5 적절한 표본수 확보하기

실용임상시험에서는 더 다양하고 넓은 범위의 환자들을 모집하기 때문에 전통적 임상시험보다 더 많은 표본수를 필요로 한다. 따라서 약물을 복용하고 있는 환자에서 치료법의 효과가 최대로 나타나지 않을 가능성도 있으며, 환자의 다양성은 오히려 치료의 효과를 정확하게 평가하기 어렵게 할 수 있다. 그러나 이러한 요소들이 실용임상시험의 타당성과 신뢰도를 위협하지는 않는다. 추가적으로 장기간 추적을 하고 싶다면 환자들의 탈락을 고려하여 더 많은 표본수를 확보하는 것이 필요하다.

1.2.2.6 환자모집 및 무작위배정

실용임상시험에서는 실제 임상적 상황에서 필요시 환자 추천이나 의뢰(referrals)를 하게 된다. 일차 의료에서 해당 진료의 중재가 어떤 잠재적 역할을 하는지 설명해야 하고, 일반의들과 상의하여 환자 전원을 어떻게 활용할지 계획을 세워야 하며, 이 경우 연구자는 일반의를 통해 환자의뢰가 실행가능하고 의미가 있는지 자문을 받아야 한다. 그리고 처음부터 되도록 동질한 두 개의 그룹으로 무작위배정을 실시하는 것이 중요하다.

1.2.2.7 결과지표의 선정

실용임상시험에서 실제 생활과 연관된 일차결과지표(primary outcome)를 선택해야 한다. 예를 들면, 환자의 기능이나 삶의 질 등이다. 또한 실용임상시험은 환자들이나 정책결정자들이 해당 이득(benefit)이 지속되는지 여부에 관심을 갖고 있기 때문에 장기간 추적관찰을 포함시키는 것이 적절하다.

1.2.2.8 분석방법

실용임상시험의 분석은 배정된 대로 분석(intention-to-treat)하는 것을 기본으로 한다. 그러나 실용임상시험에서는 환자들이 치료법을 바꿀 수 있기 때문에 치료효과가 상당부분 희석될 가능성이 있으므로 실제 상황에서 일어나는 것을 반영한 연구설계를 해야 한다. 인위적으로 치료 선택을 제한하여 자연스럽게 일어나는 것을 막기 보다는 이러한 치료의 변동들을 정확하게 기록에 남기고 보고하는 것이 중요하다. 생태학적인 요인을 바탕으로 치료법 선택 후 그 결정이 타당하다는 것을 증

명해야 하며, 환자들이 그들의 치료법을 변경하겠다는 의사결정을 존중해줄 수 있어야 하는데, 장기적 추적관찰 시에는 특히 치료법 선택과 변경에 대한 명확한 사유를 기록해 놓아야 한다.

1.2.2.9 보고와 확산

일반적으로 임상시험 보고 시 CONSORT 지침을 참고하여 보고의 수준을 높일 수 있으며, 특정 치료술에 대한 가이드라인이나, 침술의 경우 STRICTA(Standards for reporting randomized controlled trials for acupuncture)지침을 활용할 수 있다. 여기서 중요한 것은 실용임상시험 시 치료법의 모든 측면을 상세하게 보고하는 것이 필요하며, 이로 인해 연구에서 쓰여진 중재가 실제 임상현장에서 반복될 수 있도록 자세하게 기록되어야 한다.

Note 실용임상시험의 한계

실용임상시험에서는 치료과정에서 어떤 요소들이 어떠한 이익을 가져왔는지를 정확하게 파악하는 것은 어려우며, 하나의 진료 패키지가 평가의 대상이 된다. 실용임상시험의 잠재적인 문제는 표본수 증가로 인한 비용 증가 등 자원이 많이 필요하다는 것이다. 또한 실용임상시험은 눈가림(blinding)의 부족에 대하여 비판을 받을 수 있다. 전통적 임상시험의 경우 눈가림을 통해 바이어스를 줄이고, 내적 타당도를 최대화 한다. 그러나 눈가림은 생태학적으로 적절한 치료법을 선택할 수 없게 하고, 보완적 치료법에는 현실적이지 않고, 부적절할 수 있다. 따라서 실용임상시험의 경우 눈가림의 부족으로 내적 타당도가 감소하겠지만, 상대적으로 더 높은 외적 타당도를 얻을 수 있고 일반 임상 현장에서의 일반화 가능성이 높아지게 된다.

1.2.3. 사례

Smelt 등(2012)은 실용임상시험을 통해 삶의 질과 생산성을 떨어뜨릴 수 있는 편두통을 가진 환자들을 대상으로 일반 의료진이 교육 중재를 했을 때 두통과 관련 비용의 감소가 있는지 연구하였다. 이 연구는 2007년에서 2009년 사이에 네덜란드의 중소도시에 있는 64개의 일반 진료실을 군집 무작위 배정을 하였다. 일반의에게 일차진료 시 편두통 치료의 질을 향상시키기 위한 임상시험에 참여할지 의사를 물어보았고, 의사들에게 대조군에서 참여자의 행동 변화를 지양하기 위하여 중재에 관한 정보는 최소한으로 제공하였다. 환자는 18세 이상으로 제한하여 일반의의 상담 아래 연구자가 전자의무기록을 통해 선택하였다. 환자는 진료자료를 사용하여 선택

되었으며, 환자들은 6개월 이전에 triptan을 12개 이상, 12개월 이전에 24개 이상 처방을 받았을 경우 연구대상자 포함기준에 맞는 것으로 정의하였다. 군발성 두통(cluster headache), 인지결함(cognitive impairment), 심각한 정신 장애, 불치병이 있거나 네덜란드어 의사소통이 어려울 경우는 제외되었다. 포함기준에 맞는 환자를 선택한 뒤 컴퓨터로 생성된 리스트를 활용하여 진료단위를 무작위배정화 하였다. 무작위배정은 진료단위의 특성을 모르는 통계전문가가 하였고, 모든 진료단위에서 triptan을 2개 이상 사용하는 환자인구의 평균비율에 따라 층화하였다(예, 한 달에 2개 이상 triptan을 사용하는 인구가 5.9%이상인 진료단위와 5.9% 미만인 진료단위). 490명의 환자를 모집하여 233명은 중재군에, 257명은 대조군에 배정하였으며, 대조군에 속해있는 의사들은 기존의 진료를 계속하였다. 233명의 환자 중 192명(82.4%)은 편두통 치료를 평가하는 상담을 받았고, 이 환자 중 43명(22.3%)은 예방법(prophylaxis)을 시작하였다. 중재군과 대조군의 6개월간 두통 영향 지수(headache impact test at six months, HIT-6)의 변화는 통계적으로 유의한 차이가 없었다. 또한 Kessler Psychological Distress Scale(K10)을 통해 정신적 스트레스가 중재의 효과변경인자(effect modifier)인지를 분석하였다. 그 결과, 심리적 스트레스가 낮은 환자는 교육중재의 효과가 유의하게 나왔고, 심리적 스트레스가 높은 환자들은 대조군과 중재군 교육효과의 차이가 유의하지 않았다. 결론적으로 이 연구는 심리적 스트레스가 교육중재의 중요한 교란요인이 될 수 있음을 밝혔다.

1.3. 적응 임상시험

1.3.1. 개요

전통적인 RCT의 전형적인 특징은 연구 설계와 분석의 모든 측면이 사전에 정해져 있으며 임상시험 기간 동안 바뀌지 않는다는 것이다. 따라서 RCT 설계에서는 최종 결과가 나오기 이전에 그 때까지 축적된 근거를 기준으로 조기에 성과를 확인하거나 임상시험이 진행되는 동안 다른 의료기술을 추가로 고려하는 것이 어렵다(Sox and Goodman, 2012). 반면, 적응임상시험(adaptive trial)이란 임상시험을 시작한 이후에 그것의 타당성과 무결성의 약화 없이 시험자체나 통계처리과정에서 조정(adaptation)을 허용하는 임상시험 설계방법의 하나이다(Chow 등, 2005; Gallo 등, 2006; Chow and Chang, 2008). 이 연구 설계는 중간분석단계에서 축적된 자

료를 기초로 치료군의 수, 약물용량, 표본 크기, 심지어 성과지표 등 연구 설계 내용을 수정함으로써(CMTP, 2012) 유연성과 효율성을 높이고자 한다(Chow and Chang, 2008).

Note 타당성과 무결성

타당성 – 정확한 통계적 추론(보정된 p-value, 불편 추정량, 보정된 신뢰구간 등)을 제공하고, 연구의 각기 다른 단계사이의 일관성을 유지하고, 바이어스를 최소화하는 것을 의미한다.

무결성 – 확신할 수 있는 결과를 제공하고, 가능한 사전에 계획하고, 중간 분석결과에 대하여 눈가림을 유지하는 것을 의미한다.

그러나 적응임상시험은 연구를 진행하는 동안, 연구자에게 유리한 결과를 유도하기 위하여 연구 설계 내용을 유연하게 변형하는 연구가 아니다. 적응임상시험은 프로토콜 상에 중간분석시점 및 분석과정에 대해 명시하는 등 전향적으로 사전에 계획해야 한다. 조정 여부를 결정할 때는 다른 외부적인 정보들로부터 판단하는 것이 아니라 중간분석 시점까지 축적된 자료들만을 기반으로 해야 하며, 이때의 조정의 범주나 방법들도 사전 계획에 의해서만 이루어질 수 있다. 따라서 임상시험에서 적응임상시험을 적용하게 되면 그 연구의 효율성과 유연성은 증진시킬 수 있으나 임상시험의 질 자체를 바꿀 수는 없다(Gallo 등, 2006).

임상시험에서 적응임상시험설계를 적용함으로써 가질 수 있는 장점은 여러 가지가 있다. 우선, 새로운 임상적인 개발과정을 좀 더 효율적으로 가능하게 해준다. 이 연구 설계를 이용하면 좀 더 정확하게 용량 선택을 가능하게 함으로써, 임상개발의 최종적인 성공가능성을 증가시킨다. 또한 성공가능성이 낮은 시험의 경우 조기에 종료함으로써 관련 비용들을 감소시킬 수 있다. 게다가 임상시험의 여러 단계를 결합시킨 형태를 적용하게 되는 경우, 개발과정에 소요되는 시간을 줄일 수 있기도 하다. 이러한 시간이나 비용적인 측면 외에도 환자에 대한 윤리적인 측면에 있어서도 장점을 가질 수 있다. 중간분석 결과를 토대로 효과적이지 않은 그룹을 미리 탈락시킴으로써 이 군에 배정받은 환자에게 가해지는 위해를 줄일 수 있게 한다. 간단히 말해 적응임상시험방법은 기존의 전형적인 임상시험방법에 비해서 환자가 위해에 노출되는 것을 줄이고, 신약개발에 필요한 시간 및 비용을 단축시켜(CMTP, 2012), 임상적, 경제적인 편익을 얻기 위해 고안된 방법이라 할 수 있다(van der Graaf 등, 2012).

일부 사례에 의하면 적응임상시험은 전통적인 임상시험에 비해서 사회적 또는 과

학적인 가치가 떨어질 수도 있다. 적응임상시험에서는 중재결과에 대해 여러 가지를 조정할 수 있어서 잠재적으로 그 연구의 일차 연구목적 이외에 다른 시험을 하는 것이 어렵거나 불가능할 수 있다. 또한 적응임상시험이 효율성에 대한 최대한의 이점을 갖기 위해서는 빠른 시간 내에 자료 수집을 필요로 하는데, 이것들은 전반적인 임상시험에 걸친 시간에 비해 짧은 추적 기간 안에 중재결과를 확인할 수 있을 때 가능하다(Gallo 등, 2006). 이를 다르게 해석하자면 이것은 검증된 대리지표(surrogate marker)가 없는 한 장기추적 임상시험(long-term survival trial)에서는 적합하지 않다는 의미가 된다. 다른 문제로 이러한 적응임상시험이 재현될 수 있냐는 것인데, 이로 인해 결과의 일반화와 재생산 측면에서 제한될 수밖에 없다(van der Graaf 등, 2012).

적응임상시험연구에서 베이지안 방법(Bayesian approach)과 전통적 통계방법(frequentist approach)이 모두 사용가능하나, 베이지안 방법은 특히나 적응임상시험연구에 적합하다. 베이지안 방법은 전통적인 통계방법과는 다르게 축적된 증거(예. 중간분석결과)와 사전정보(prior information) (예. 기존 관련 문헌의 자료) 모두를 이용할 수 있다(Luce 등, 2009). 이 사전정보는 현재의 자료들과 결합하여 이를 이용하여 결론을 도출할 수 있다(CMTP, 2012).

Note 베이지안 방법

베이지안 방법은 주어진 사전정보 또는 전문가 의견을 바탕으로 사전분포(prior distribution)로 계량화하고, 이 사전분포를 자료와 결합하여 사후분포(posterior distribution)를 구함으로써 치료효과 등의 관심 모수에 대해 추정하는 통계적 방법이다. 베이지안 방법을 적용하여 치료효과를 추정할 경우에는 추정된 결과를 확률적 표현으로 제시할 수 있다는 장점 등이 있기 때문에 최근 적응임상시험에서 유용하게 적용되고 있다(Pullenayegum, 2012).

1.3.2. 방법

1.3.2.1. 적용범위 및 유형

적응임상시험은 다음의 요인들에 대해 조정을 허용한다.

- 임상시험의 목적
- 임상시험에서 고려되는 구체적 연구 가설
- 연구가설에서 쓰이는 임상결과 측정변수

- 임상시험에 행해지는 조건
- 모집단/연구대상자
- 환자의 실험군 할당
- 무작위배정 과정
- 약물 용량이나 빈도와 같은 치료 요법
- 예상되는 통계적 분석 방법

적응임상시험에서 한 가지 조정은 단순하게 그 범주 내에서의 조정만을 의미하지는 않는다. 예를 들어 연구가설이 변경되는 경우, 그로 인해 다른 것들이 모두 영향을 받아 변형이 일어나게 되기 때문이다. 조정이 가능한 범주에 따라 적응임상시험 연구 설계 방법을 다음과 같이 나눌 수 있으나 이 유형들이 여러 가지 접목된 형태로 이루어지고 있다.

(1) 가설 조정 설계

가설 조정 설계(hypothesis-adaptive design)는 중간분석 결과를 기초로 가설의 변경이나 수정을 가능하게 하는 연구 설계방법이다. 예를 들어 우위성 가설(superiority hypothesis)에서 비열등성 가설(non-inferiority hypothesis)로의 전환, 연구의 일차 결과지표와 이차 결과지표 사이의 전환 등이 있다. 우위성 가설에서 비열등성 가설로의 전환에서, 비열등성 허용 한계(non-inferiority margin)의 선택은 원하는 수준의 검정력을 얻기 위한 표본 수 조정 및 통계적 검정 방법 등에 중대한 영향을 미치므로 주의하여 결정해야 한다.

(2) 표본수 재추정 설계

표본수 재추정 설계(sample size re-estimation design)는 중간분석에서 관찰된 자료를 기반으로 표본 수를 조정하거나 재추정하는 연구 설계방법이다(Gallo 등, 2006). 그러나 표본수 재추정 설계 역시 중간에 관찰된 자료를 기초로 하기 때문에, 추정치를 토대로 산출하여 바이어스된 표본 수 산출이 일어날 수 있는 기존 설계의 한계를 여전히 갖게 된다. 게다가 적은 수의 환자를 대상으로 시작하여 표본수를 재추정하기 때문에 임상적으로나 통계적으로 좋은 방법이 아닐 수 있다. 따라서 이 연구 설계 방법의 경우 제한된 수의 대상에서 관찰된 차이를 기반으로 하기 때문에 그 결과가 바이어스 되거나 왜곡될 수 있음을 주의해야 한다.

(3) 치료 전환 설계

치료 전환 설계(adaptive treatment-switching design)는 연구자가 환자의 치료에 대한 효능 또는 안전성의 부족을 근거로 임상시험 초기에 배정된 치료군을 변경할 수 있도록 하는 연구 설계방법이다. 윤리적인 문제측면으로는 예를 들어 종양학 연구와 같은 경우에 환자의 생존문제를 위해서도 가능하다. 그런데 이렇게 치료군을 바꾼 환자가 많은 경우, 검정하려던 가설이 변할 수 있음을 주의해야하며, 이 경우 원하는 수준의 검정력을 확보하기 위해서는 표본수 조정이 필요하다.

(4) 무작위 배정 조정 설계

무작위 배정 조정 설계(adaptive randomization design)는 성공확률을 증가시키기 위해서 치료군 배정 확률을 다르게 하는 등 무작위 배정 방법을 변화시키는 것을 허용하는 연구 설계 방법이다. 이 설계는 치료방법, 공변량(covariate), 반응 등에 무작위 배정 방법을 조정하는 것을 포함한다. 하지만 이 연구 설계방법은 임상시험의 성공확률을 증가시킴에도 불구하고, 이전 대상자의 반응에 의존하여 다음 대상자에 무작위배정을 진행하기 때문에 장기적 치료기간이 필요하거나 규모가 큰 임상시험에서는 실현하기 힘들다. 또한 치료효과에 대한 통계적 추정이 힘든 경우도 있으므로 주의해야 한다.

(5) 그룹 축차 설계

그룹 축차 설계(group sequential design)는 중간분석(interim analysis) 결과를 기초로 프로토콜을 개정하거나 임상시험의 안전성과 유효성 유무에 따라 임상시험을 완료하기 이전에 중단하는 것을 가능하게 하는 연구 설계방법을 말한다. 하지만 이 연구 설계의 경우 프로토콜의 개정이나 추가적인 조정으로 인해 대상 환자군의 이동이 일어난다면, 제1종 오류를 통제할 수 없어서 적합하지 않을 수 있다. 따라서 이 연구 설계 방법을 적용하는 경우 제1종 오류를 보정하기 위한 노력이 필요하다.

(6) 열등 반응 탈락 설계

열등 반응 탈락 설계(drop-the-loser design)는 약의 용량에 대해 불확실성이 있는 2상 임상시험에서 유용한 것으로, 열등한 치료군의 탈락이나 치료군을 추가하는 것을 허용하는 방법이다. 열등 반응 탈락 설계에서 열등한 환자군을 정하는 선택기준과 이 군의 탈락여부를 결정하는 방법은 중요하다. 전형적으로 2가지 단계로 나눌 수 있는데, 첫 단계가 끝날 때에 사전에 정해진 범주에 따라 이전의 치료군을 탈락시킨 후, 남은 치료반응을 보인 군들에 대해서만 다음 단계로 진행하게 된다.

실제로, 연구는 두 번째 단계(연구의 마지막)까지 수행하면 원하는 수준의 검정력을 얻게 되는데, 이는 다르게 말하자면 첫 단계에서는 열등한 치료군을 탈락시킬 정도의 통계적인 검정력이 없다는 것이 된다. 따라서 통계적인 검정력에 대한 주의가 필요하다.

(7) 약물 용량 조절 설계

일반적으로 초기 임상시험에서는 다음 임상시험 단계로 진행하기 위해 약물용량을 정할 때 약물의 최소유효량(Minimum Effective Dose, MED)이나 최대내성용량(Maximum Tolerable Dose, MTD)을 정하게 된다. 전형적인 용량 확정 설계(dose finding design)는 몇 가지 고정된 용량의 평행연구를 통해서 진행되어 큰 표본수가 필요하다. 실제의 용량-반응 곡선 하에 정확한 선택이 된 경우를 제외하면, 연구자들은 높은 수준의 정보에 기초하지 않은 채 약물의 용량을 선택하곤 한다. 반면에 약물 용량 조절 설계(adaptive dose finding design)를 사용하게 되면, 전반적인 과정에 대하여 효율적으로 용량-반응의 특성을 볼 수 있다. 비효율적이거나 안전하지 않은 용량에 대해서는 치료군을 줄이고, 더 효과적인 용량에 대해서 치료군을 증가시키게 된다.

베이지안 방법을 접목한 연속적 재평가 방법(Continual Re-assessment Method, CRM)이나 용량-반응 곡선을 추정하기 위해 비모수 조정 설계 방법(nonparametric adaptive urn design approach)을 사용한다(O'Quigley and Shen, 1996; Chang and Chow, 2005)

(8) 2상/3상 통합 임상시험 설계

임상시험에서 전형적인 접근법의 경우 2상에서는 확실하고 가치 있는 정보를 얻고, 3상에서는 연구에 대한 검정력을 더하는 것이 목적이다. 그러나 이런 전형적인 접근법은 타당성과 효율성이 떨어질 수 있다는 문제점을 토대로 통합 설계가 대두되었다. 2상/3상 통합 임상시험 설계(adaptive seamless phase II/III trial design)는 의도하고자 하는 임상시험의 목적을 달성하기 위해, 원래는 후기 2상(phase IIb)과 3상(phase III)에서 별도로 수행해야 하는 과정을 하나의 프로토콜로 이루어진 단일 연구에서 얻을 수 있도록 조정한 연구방법을 말한다. 이는 두 단계로, 학습단계(learning stage)와 확인단계(confirmatory stage)를 포함한다. 또한 조정 전후에 상관없이 환자의 모든 자료를 사용하여 최종분석을 한다. 단, 종합적인 분석과 각 임상연구 단계에서의 분석 결과가 다른 의미를 가질 때 어떻게 해석해야 하는지에 관해서 여전히 더 많은 연구가 필요한 실정이다.

(9) 다중 조정 설계

다중 조정 설계(multiple adaptive design)는 위의 언급했던 여러 연구 설계방법들을 서로 접목한 형태의 연구 설계이다. 그러나 실제로, 이러한 다중 조정 설계의 경우 통계적인 추론이 어렵기 때문에 계획단계에서 이 연구 설계의 수행에 대한 충분한 평가와 모의실험이 필요하다.

1.3.2.2. 주의사항

적응임상시험은 여러 면에서 유용하지만 오남용 될 소지가 많기 때문에 적용하는데 있어서 다음과 같은 사항에 관하여 주의가 필요하다.

(1) 눈가림의 유지는 옳은가?

적응임상시험에 대해서는 타당성과 무결성이 모두 유지될 수 없다는 부정적인 의견이 있다. 예를 들어 연구 설계를 조정하기 위해서는 중간 결과에 대한 분석이 있어야 하는데, 이런 중간분석 결과는 안전성 모니터링 위원회, 연구자, 그리고 참가자들에게 영향을 줄 수 있다. 이러한 중간분석 결과가 공개되었을 때, 연구자들이 다음에 참가자를 배정할 때 더 유리한 결과가 나올 수 있도록 인위적으로 배정할 수도 있기 때문에 시험과정에 있어서 바이어스를 유발할 수 있다(van der Graaf 등, 2012). 즉, 적응임상시험을 진행하는데 있어서 필수적인 단계이기도 한 중간분석의 결과들을 해석하고 처리하는 과정에서 눈가림의 유지 여부는 중요한 고려대상이 된다. 중간분석 결과의 눈가림을 해제해야 하는지 혹은 유지해야 하는지에 따라서는 여러 논란이 있으며, 어떤 것이 더 우월한 방법인지에 대해서는 이견이 존재한다.

예를 들어 중간분석에서 관찰된 자료를 이용하여 표본수를 다시 산출하는 연구 설계인 표본수 재추정 설계는 작은 수의 표본에서 얻어진 추정치를 마치 실제 값인 것처럼 이용하여 계산하기 때문에 부정확성, 바이어스가 나타나거나 또는 결론이 왜곡될 수 있다는 문제가 있다. 이때 눈가림을 유지한 상황에서 표본수를 재추정하게 되면, 내부연구진으로도 다음 과정을 진행할 수 있고 제1종 오류에 대한 영향을 최소화 시킬 수 있다. 단 눈가림이 유지된 채 중간분석 결과를 해석하게 되면, 장애모수(nuisance parameter)의 추정은 잘못될 수 있고, 이로 인해 정확하지 않은 재보정이 일어나게 된다. 반면에 눈가림을 해제하고 중간분석 결과를 해석하게 되는 경우, 더 정확한 표본수를 추정할 수 있게 해준다. 단, 내부 연구진과는 무관한 외부의 연구진을 구성하여 진행해야 한다는 단점이 있다.

(2) 자료 분석 시 통계적인 처리와 제1종 오류에 대한 통제는 어떻게 해야 하는가?

적응임상시험의 결과를 임상적 또는 정책적으로 해석하기 위해서는 통계적인 접근법을 이용한 가설 검정이 필수적이다. 이론적으로, 임상시험에 대한 조정 측면에 있어서 제1종 오류를 조절하면서 얻어진 자료에 대해서 최대한의 유연성을 허용하는 것은 가능하다. 그러나 최대한 조정이 가능한 분야와 조정 범위를 확실하게 하는 것은 필요하다. 이것은 결과에 대한 해석을 쉽게 하고, 적응임상시험으로부터 나온 추론을 쉽게 받아드릴 수 있도록 돕기 때문이다(Gallo 등, 2006).

전통적 통계방법에서 적응임상시험법에 관한 치료 효과에 대한 추정과 신뢰구간 설정에 대한 이슈는 아직 완벽히 해결되지 않았다. 자료 분석 시에 통계적 추론의 정확도와 신뢰도에 영향을 미칠만한 바이어스가 생길 수 있다는 것을 고려해야 하며, 미리 선정한 유의수준에 따른 제1종 오류를 조절할 수 있는지 역시 중요한 고려사항이다.

적응임상시험에서는 중간분석이라는 단계가 필요하며 이러한 중간분석은 1회 혹은 그 이상일 수 있다. 이러한 중간분석이 여러 차례 이루어지면서 통계적 가설검정을 여러 번 하는 다중검정이 이루어지게 된다. 이러한 다중검정을 할 때, 개별가설검정을 5%의 유의수준에서 검정한다면 모든 귀무가설이 참임에도 불구하고, 적어도 하나의 귀무가설을 기각하게 되는 오류가 5%보다 커지게 되는 다중성의 문제가 생기게 된다. 이러한 다중성의 문제를 해결하지 않고 5% 유의수준으로 검정하는 중간분석을 20회 한다면 평균적으로 1번은 효과가 없음에도 불구하고 효과가 있다고 결론지어지도록 귀무가설을 기각하게 되는 모순이 발생하게 된다. 따라서 이 경우, 이러한 다중성의 문제를 어떻게 보정할지에 대해서 미리 계획서상에서 다룰 필요가 있다. 다중검정으로 인해 발생하는 문제를 해결하기 위한 통계방법들로 Pocock 방법, O'Brien & Fleming 방법 등이 있다. 따라서 다중검정을 하게 되는 경우, 이러한 제1종 오류를 통제할 수 있는 적당한 방법을 이용하여 통계처리를 해야 한다(강승호, 2010).

(3) 독립적인 자료검토위원회는 필요한가 또 그 역할은 무엇인가?

적응임상시험에서 조정이 이루어진 경우 전반적인 임상시험과정이 독립적인 자료검토위원회(Data Monitoring Committee, DMC)에 의해서 검토되어야 결과의 해석이 이해 당사자(스폰서, 연구자 등)로부터 차단되기 때문에 임상시험의 무결성과 결과의 안전성이 보장될 수 있다(U.S. FDA, 2006). 이들의 자료 검토는 엄격히 기밀사항이고, 자료검토위원회에서 운영위원회와 스폰서로의 보고 과정은 미리 정해져 있다. 중간결과에 대한 접근이 허용된 이해당사자들은 모든 임상시험 과정에 직

접적으로 참여해서는 안 된다. 또한 정보가 공개되었을 때 나타날 수 있는 부정적인 영향을 미리 숙지하고 있어야 한다. 특히 자료의 공개는 제일 적은 수의 스폰서 대표자들에게 미리 정해진 시간에 한하여 최소한의 정보만이 공개되는 등의 방법으로 중간 자료에 대한 접근을 최소화 시켜야한다(Gallo 등, 2006). 자료검토위원회는 임상시험이 안전성과 유효성 측면에서 이점이 없을 때 시험을 중단할 수 있는 규칙과 분석일정을 정해야 하며, 이에 근거하여 모니터링 해야 한다(Chow and Chang, 2008).

(4) 윤리적인 측면에서 시험 참가자 선택시 문제는 무엇인가?

도덕적으로 임상시험이 인정받기 위해서는 시험 참가자들은 공정하게 선택되어야 한다. 적응임상시험의 목표는 참가자들을 불필요한 치료법에 노출을 최소한으로 하는 것이고, 참가자들에게 최대로 유익한 치료를 받게 하는 것이다. 하지만 나중에 참가하는 참가자들이 좀 더 이익을 받게 되는데, 그렇게 되면 실험이 진행됨에 따라 환자들에게 더 유익한 치료를 제공하지 않는데 있어서 임상시험에 참가한 의사들에게 윤리적인 문제가 있을 수 있다. 연구 설계의 이해에 따라서 이익의 정도가 갈린다면 이것은 전혀 공평하지 않은 것이다. 예를 들어, 효과적이지 못한 용량의 치료를 받을 때 환자들은 임상시험에서 곧 제외될 수 있고, 연구 후반에 참가 할수록 좋은 치료를 받는다는 것을 모르는 참가자 있다면 이것은 불공평한 것이 된다. 적응임상시험 참가자들의 부담은 일반적인 시험보다 더 높을 수도 있다. 비록 전반적인 참가자의 수가 비교적 적기는 하지만, 중간 분석을 위해서 다른 때보다 좀 더 길게 시험에 참가해야할 수도 있기 때문이다. 그리고 치료반응에 대한 무작위배정을 조정한 경우 실험 참가자들에 있어서의 이익이 평소수준이거나 오히려 없을 수도 있다는 의견도 있다(van der Graaf 등, 2012).

1.3.3. 사례

1.3.3.1. 약물 용량 조절 설계

Lewis 등(2011)은 절뚝거림(claudication)의 치료제로써 포스포디에스테라아제(phosphodiesterase, PDE) 억제제로 개발된 K-134의 최대내성용량을 찾기 위한 비교임상시험에서 적응임상시험 설계 방법을 활용하였다. PDE 억제제는 말초동맥질환(peripheral artery disease, PAD)에 의한 절뚝거림 환자에서 걷는 능력의 향상을 보여주는데, K-134는 선택적 PDE3 억제제로 이 질환의 잠재적인 치료제로써

개발되었다. 일반적으로 PAD 환자들에서 K-134와 다른 PDE3 억제제의 병용은 허혈성 심질환, 빈맥, 저혈압을 유발하는 등 안전성과 내성의 문제를 일으키는 것으로 알려져 있다.

이 연구는 안정적이나 간헐적인 절뚝거림(stable intermittent claudication)질환의 치료제로써 K-134가 작용하는가에 대한 평가목적으로 설계된 제 2상 임상시험이며, 미국과 러시아에서 수행된 위약과 활성약물을 3가지 용량(25 mg, 50 mg, 100 mg) 에 대하여 비교하는 형태의 이중눈가림, 무작위 배정 연구이다.

이 연구의 주요 목적은 26주간의 치료이후에 K-134 의 최대내성용량의 위약과의 비교이다. 조정에 대한 의사결정을 위한 자료를 제공하기 위해서 연구기간의 초기에 안전성 측정을 포함하고 있다. 이 연구는 연구 기간 내에 관찰되는 부작용을 기초로 안전하다고 받아들여지는 최대 용량을 효과적으로 규명하기 위해서 적응임상시험 전략을 사용하였다.

연구계획서에 안전성과 내성의 결과지표에 대한 기준을 미리 규명해놓았고, 최대표적 표본 수는 각 군당 85명이었다. 199명을 무작위배정하고 143명으로부터 28일간의 자료를 중간 분석한 결과, 자료검토위원회는 가장 적은 용량(25 mg)을 투여하는 치료군에 대해서 중단할 것을 추천하였다. 만약 25 mg K-134 치료군이 탈락되지 않고 모든 군에 대해서 연구를 끝까지 진행했다면 약 43명의 추가 환자가 필요하였을 것으로 추정되었다. 즉, 적응임상시험방법을 적용함으로써 효율성을 개선시킬 수 있었다.

1.3.3.2. 2상/3상 통합 임상시험 설계

Levin 등(2011)은 급성 허혈성 뇌졸중 치료제인 알티피에이(r-tPA)와 테넥테플라제(tenecteplase)의 비교임상시험에서 2상/3상 통합 임상시험 설계를 활용하였다. 이 연구는 시험 설계와 포괄적인 제 1종 오류를 검토하고, 규제당국의 승인과 관련한 이슈에 관해 토의하고, 미래의 유사한 시험 설계의 승인을 촉진하는데 그 목적이 있다. 임상시험 2상의 초기단계에서 조기의 중재결과를 토대로 테넥테플라제의 3가지 용량(0.10, 0.25, 0.40 mg/kg)중에서 3상에서 사용할 한 가지 용량을 선택하는 과정에서 열등한 효과를 보이는 치료군을 순차적으로 제외하는 그룹 축차설계를 적용하였다. 또한 2상이 끝나는 시점에서, 임상적인 효능과 안전성을 이전에 선택된 용량의 테넥테플라제를 투여 받은 환자군과 알티피에이를 투여 받은 환자군을 비교하였다. 이 비교를 통해 효과 없음을 근거로 2상에서 종료할지 3상으로 지속할지 여부를 결정하였다. 연구가 3상으로 지속된다고 가정하였을 때, 3상의 첫

번째 중간분석은 2상의 마지막단계에서의 결정으로 갈음하여 사용함으로써 통합 설계를 활용하였다.

이 연구는 조기종료가 없다고 가정했을 때, 4번의 중간분석과 1번의 최종 분석에서 1,908명의 환자를 대상으로 하고, 실험별 제1종 오류(experiment-wise type I error)는 0.05 이하로 설정하였다. 1,000개 이상의 시나리오에 대해서 각 시나리오당 40,000번을 반복했을 때, 3상에서 제1종 오류의 최대치는 0.038이다. 이 실험방법은 여러 차례의 조정을 적용함에 따라 제1종 오류가 증가되는 문제가 발생할 수 있다. 그러나 용량 선택과정에서의 제 1종 오류의 증가는 검정통계량의 연속성 수정(continuity correction)의 0.5배 이하로 상쇄되고, 반복되는 중간분석으로 인한 제1종 오류의 증가는 첫 중간분석에서 효과 없음을 이유로 임상시험이 중단되는 것으로부터 상쇄될 수 있다.

이 임상시험은 복잡한 시험 설계와 진화하는 규제 요구조건들로 인해 검토 과정이 연장된다는 한계점을 갖지만, 시험 설계방법은 여러 가지로 혁신적이고 효율적이다. 이러한 적응임상시험이 적용되기 위해서는 각 프로토콜에서 3상 가설 검정을 위한 제1종 오류는 잘 통제되어야 하고, 경험적으로 계산되어야 한다. 또한 처음 시험 설계 단계에서부터 규제 검토자들과의 의사소통을 위한 시간을 필요로 한다. 제1종 오류 통제와 관련하여 프로토콜을 위반한 경우의 제1종 오류의 논증을 포함할지 여부와 제1종 오류 통제의 모의실험을 수용할지 여부에 대해서도 분명히 해야 할 필요가 있다.

참고문헌

강승호, 신약개발에 필요한 의학통계학, 자유아카데미, 2010.

Agency for Healthcare Research and Quality. Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews, Version 1.0 [Draft posted Oct. 2007]. Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf

Boden WE, O'Rourke RA, Teo KK, et al, for the COURAGE Trial Research Group, for the COURAGE Trial Coprincipal Investigators and Study Coordinators. Optimal medical therapy with or without PCI for stable coronary disease. N Engl J Med. 2007;356:1503–1516.

Boden WE, O'Rourke RA, Teo KK, et al. Design and rationale of the Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation (COURAGE) trial: Veterans Affairs Cooperative Studies Program no. 424. Am Heart J. 2006;151:1173–1179.

Bryan R. Luce, Judith M. Kramer, Steven N. Goodman, Jason T. Connor, Sean Tunis, Danielle Whicher, J. Sanford Schwartz; Rethinking Randomized Clinical Trials for Comparative Effectiveness Research: The Need for Transformational Change. Annals of Internal Medicine. 2009 Aug;151(3):206–209.

Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomized trials. BMJ. 2004; 328: 702–708.

Campbell MK, Grimshaw JM. Cluster randomized trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. BMJ. 1998; 317: 1171–1172.

Center for Medical Technology Policy, Making Informed Decisions: Assessing the Strengths and Weaknesses of Study Designs and Analytic Methods for Comparative Effectiveness Research, 2012

Chang M, Chow SC. A hybrid Bayesian adaptive design for dose response trials. J Biopharm Stat. 2005;15(4):677–691.

Chow SC, Chang M, Pong A. Statistical consideration of adaptive methods in clinical development. J Biopharm Stat. 2005;15(4):575–591.

Chow SC, Chang M. Adaptive design methods in clinical trials – a review. Orphanet J Rare Dis. 2008;3:11.

Detre KM, Peduzzi P, Takaro T, Hultgren H, Murphy M, Kroncke G,for the Veterans

Administration Coronary Artery Bypass Surgery Cooperative Study Group. Eleven-year survival in the Veterans Administration Randomized Trial of Coronary Artery Bypass Surgery for Stable Angina. N Engl J Med. 1984;311:1333-1339.

Donner A. Some aspects of the design and analysis of cluster randomized trials. J Roy Stat Soc C-App. 1998; 47: 95-113.

Eldridge S, Ashby D, Bennett C, et al. International and external validity of cluster randomized trials: systematic review of recent trials. BMJ. 2008; 336: 876-880.

Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development--an Executive Summary of the PhRMA Working Group. J Biopharm Stat. May 2006;16(3):275-283; discussion 285-291, 293-278, 311-272.

Glynn RJ, Brookhart MA, Stedman M, et al. Design of Cluster-randomized trials of quality improvement interventions aimed at medical care providers. Medical Care. 2007; 45: S38-S43.

Hutton JL. Are distinctive ethical principles required for cluster randomized trials? Stat Med. 2001; 20: 473-488.

Levin B, Thompson JL, Chakraborty B, Levy G, MacArthur R, Haley EC. Statistical aspects of the TNK-S2B trial of tenecteplase versus alteplase in acute ischemic stroke: an efficient, dose-adaptive, seamless phase II/III design. Clin Trials. Aug 2011;8(4):398-407.

Lewis RJ, Connor JT, Teerlink JR, et al. Application of adaptive design and decision making to a phase II trial of a phosphodiesterase inhibitor for the treatment of intermittent claudication. Trials. 2011;12:134.

Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. Ann Intern Med. 2009 Aug;151(3):206-209.

Mazor KM, Sabin JE, Boudreau D, et al. Cluster Randomized Trials: opportunities and barriers identified by leaders of eight health plans. Medical Care. 2007; 45: S29-S37.

O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. Biometrics.1996 Jun;52(2):673-684.

Peduzzi P, Kyriakides T, O'Connor TZ, Guarino P, Warren SR, Huang GD. Methodological issues in comparative effectiveness research: clinical trials. Am J Med. 2010 Dec;123(12 Suppl 1):e8-15.

Platt R, Takvorian SU, Septimus E, et al. Cluster ranidomized trials in comparative

effectiveness research: randomizing hospitals to test methods for prevention of healthcare-associated infections. Medical Care. 2010. 48: S52-S57.

Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. J Eval Clin Pract. 2005; 11: 479-483.

Puffer S, Torgerson DJ, Watson J. Evidence for risk of bias in cluster randomized trials: review of recent trials published in three general medical journals. BMJ. 2003; 327: 785-789.

Pullenayegum EM. Adaptive Bayesian randomized trials: realizing their potential. J Bone Joint Surg Am. 2012 Jul;94 Suppl 1:29-33.

Saag KG, Mohr PE, Esmail L, et al. Improving the efficiency and effectiveness of pragmatic clinical trials in older adults in the United States. Contemporary Clinical Trials. 2012.

Smelt AFH, Blom JW, Dekker F, Akker ME, Neven AK, Zitman FG, et al. A proactive approach to migraine in primary care: a pragmatic randomized controlled trial. CMAJ March 6 2012; 184: E224-231.

Sox HC and Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health. 2012 Apr;33:425-445.

US Food and Drug Administration. Guidance for Clinical Trial Sponsors. Establishment and Operation of Clinical Trial Data Monitoring Committees. Rockville MD: FDA; 2006. Available at: http://www.fda.gov/cber/gdlns/clindatmon.htm.

van der Graaf R, Roes KC, van Delden JJ. Adaptive trials in clinical research: scientific and ethical issues to consider. JAMA. 2012 Jun ;307(22):2379-2380.

PART 2

비교효과 관찰연구

비교효과 관찰연구(observational research)는 치료법이 연구자에 의해 결정되는 임상시험과 달리, 자연적으로 발생한 인자의 노출을 탐구하여 특정 질병과의 상관관계를 조사하는 방법이다(Concato, 2010). 주로 치료법이나 중재법을 포함한 요인의 노출 여부에 의해 구분된 독립된 두 집단의 질병 빈도를 계산하여 그 차이를 연관성 지표(measure of association)로 삼는다. 특정 요인과 질병발생과의 관련 정도를 측정할 시 필요한 가정 중의 하나는 질병의 위험이 모든 비교 그룹에서 동일해야 한다는 것이다(AHRQ, 2012).

비교효과 연구에서 임상시험에 비해 관찰연구가 갖는 이점은 다음과 같다. 첫째, 임상연구의 경우 요인의 인위적인 무작위 배정을 통해 내적타당도(internal validity)는 향상되지만, 대신 일반화가능성(generalizability)은 약화된다. 또한, 임상연구는 제한된 인구집단을 대상으로 이상적인 환경에서 치료법의 효과를 조사하는 반면, 관찰연구는 위험집단을 잘 대변해주는 연구 집단을 대상으로 보다 현실적인 환경에서 수행된다. 비교효과 연구는 특정 치료법의 모든 잠재적 사용 집단에 적용 가능한, 보다 일반적인 효과를 검증해내고자 하는 연구이므로 일반화 측면에서 이점을 가진 관찰연구가 더 적합하다(AHRQ, 2012). 둘째, 비교효과 연구에서 임상연구(즉, 무작위 배정 임상시험[RCT])가 적합하지 않을 수 있다. 예를 들어, 새로운 중재법이 기존의 치료법에 비해 확실히 효과가 있는 경우나 반대로 환자의 이익에 위배되는 경우는 임상연구를 수행할 수 없다. 또한, 확실한 치료법이 존재할 경우나 치료 약물이 이미 시중에 유통되었을 경우, 다양한 집단의 이해관계가 얽히면서 임상 연구의 수행이 어려워진다(Norris 등, 2010). 셋째, 여러 중재법들에 대해 직접 비교(head-to-head comparison)를 할 경우, 효과크기(effect size)가 작아서 상대적으로 많은 수의 환자들이나 장기간의 추적이 필요하다. 따라서 보다 신속하고 저비용으로 많은 집단을 연구할 수 있는 관찰연구가 임상연구에 비해 비교효과 연구에서 더 효율적이다. 마지막으로, 관찰연구와 임상연구를 비교한 여러 연구에 따르면, 잘 설계된 관찰연구는 임상시험과 비교 가능한 유효한 결과를 도출해낸다. 예를 들어, 폐경기 여성군에서 호르몬 대체 요법(hormone replacement therapy) 사용과 다양한 질병발생(예. 관상 동맥성 심장질환, 뇌졸중 등) 간의 상관관계를 조사한 임상시험과 관찰연구를 비교한 결과, 환자군의 차이점 (예. 사회경제적 지위 등), 치료법, 추적 기간 등이 잘 고려되었을 경우 두 연구 방법은 비교적 비슷한 결과를 도출해냈다(Grodstein, 2003). 또한, 5개의 임상적 주제에 대한 99개의 개별 연구를 메타 분석한 결과도 임상시험과 관찰연구가 비슷한 결과를 도출해냄을 보여줬다(Concato, 2010). 이는 비교효과연구에서 관찰연구가 임상시험을 대신하여 유용한 방법이 될 수 있음을 보여준다.

2.1. 관찰연구 설계

관찰연구를 설계할 때 가장 먼저 고려해야 할 사항은 연구참여 집단(study population)이 선택될 대상인구(population)와 자료원이다. 주로 일반 인구(general population)가 선호되지만, 공변량(covariates)에 대한 자료 확보와 같은 이점으로 선택된 인구(예. 특정 약물을 복용하는 집단이나 질병등록사업에 등록된 환자)가 선호되는 경우도 있다. 자료원의 선택은 연구의 특성에 의해 결정되어진다. 우선 연구하고자 하는 노출변수와 결과변수를 정의하고 노출인자의 선택에 영향을 미칠 수 있는 결과변수 예측요인, 즉 교란요인(potential confounders)을 정의한다. 이를 바탕으로, 기 존재하는 자료의 유용성 및 타당성과 새로운 자료나 추가 자료 수집의 필요성 여부를 시간과 비용의 관점에서 비교 분석한다(AHRQ, 2012). 비교효과 연구의 자료원에 대해서는 다음 단락에서 보다 자세히 기술하겠다.

비교효과 연구에서 선정과 제외 기준(inclusion and exclusion criteria)은 연구기간과 기준을 정한 날짜와 함께 명확하게 정의 내려져야 한다. 특히 주의할 점은 기준 설정 시 모든 연구집단에게 동일한 기간이 적용되어야 한다는 것이다. 만약 연구 대상자들 사이에 연구에 참여한 기간의 차이가 존재할 경우 바이어스(bias)가 유발될 수 있기 때문에 비교군 간의 기간 차이는 주의 깊게 평가되어야한다. 각각의 기준은 기초자료(baseline data)를 토대로 결정되며 연구 진행단계나 추적기간 동안 습득된 정보를 토대로 변경될 수 없다. 잘 정의된 선정 및 제외 기준은 관찰 연구의 내적타당도를 향상시킬 수 있다. 동반상병(comorbidity)이 잠재적 교란요인인 연구를 예로 들면, 대개 동반상병의 진단은 특이도(specificity)는 높지만 민감도(sensitivity)는 낮다. 이런 경우, 연구대상 집단을 동반상병 정보를 갖고 있는 집단으로 제한할 경우 교란요인을 통제할 수 있다. 하지만 이 경우에서 알 수 있듯이 선정 및 제외 기준은 일반화에 영향을 미치게 된다(AHRQ, 2012).

치료군과 비치료군을 비교할 때 치료법이 유사한 중증도의 환자에게 시행되지 않을 경우 발생하는 적응증에 의한 교란(confounding by indication)과 건강 상태에 따라 치료법의 처치가 달라짐으로써 발생하는 교란은 가장 중대한 문제점이다. 따라서 대조군의 선택은 교란효과를 제어하기 위해 매우 중요하다. 예를 들어, 치료군을 같거나 혹은 비슷한 지표(indication, 치료 선택에 영향을 주어 결과변수에까지 영향을 미치는 특성)를 가진 대조군과 비교하면 적응증에 의한 교란 등 잠재적 바이어스를 제어할 수 있다 (예. 당뇨병 환자 중 인슐린(insulin) 복용집단과 글리타존

(glitazones) 복용집단). 또한, 같은 지표를 가진 활성 대조군(active comparator) 즉, 다른 치료약을 복용하고 있는 집단)을 이용할 경우, 치료를 결정한 시점을 확인하여 모든 연구 대상자들의 추적을 동일한 시점에서 시작할 수 있다는 이점이 있다(AHRQ, 2012).

2.1.1. 비교효과 관찰연구의 유형

연구 주제에 적합하고 연구 결과의 타당성을 위해하는 요소들을 최소화하는 연구를 설계하는 것은 관찰 연구 수행 시 가장 중요한 과정이다. 설계 단계에서 결정된 사항들은 연구 결과에 영향을 미치며 이미 설계가 끝난 후에는 보정이 불가능하므로 매우 주의를 기울여야한다. 비교효과 연구에서 사용 가능한 관찰연구의 유형으로는 크게 코호트 연구, 환자-대조군 연구, 환자-코호트 연구가 있으며, 그 외에 환자-교차설계 연구, 환자-시간-대조군 연구와 같은 유형이 있다(AHRQ, 2012).

2.1.1.1. 코호트 연구

코호트 연구(cohort study)는 연구 시작 시점에서 노출 인자(exposure factor)의 노출 여부에 따라 연구 집단을 구성하고 이들을 일정 기간 동안 추적하여 특정 질병의 발생 여부를 관찰하는 연구를 지칭한다. 대개 질병의 최초 발생만을 고려하기 때문에 연구 시작 시 특정 질병을 이미 가지고 있는 코호트 구성원은 배제되어야 한다. 이때 배제기준은 기초자료를 기준으로 해야 하며, 추적기간 동안 새로이 얻게 된 정보를 기준으로 환자가 배제될 경우 바이어스가 생길 수 있다. 코호트 연구는 연구 방법에 따라 전향적과 후향적으로 나뉘며, 두 경우 모두 요인 인자 노출보다 질병 발생이 선행되어서는 안 된다.

비교효과 연구에서 코호트 연구가 갖는 장점은 첫째, 요인 노출과 질병 발생의 시간적 선후관계가 명확하고, 둘째, 실제 질병 발생률(actual incidence)을 예측할 수 있으며, 셋째, 연구 가설에 제시된 질병뿐만 아니라 여러 다양한 질병들과 주어진 치료법의 상관관계를 연구할 수 있고, 넷째, 임상시험과의 비교가 비교적 용이하다는 점이다.

이에 반해, 코호트 연구의 단점은 질병 발생률이 낮은 경우 대규모 코호트가 필요하거나 관찰 기간이 길어야 하므로 비효율적이라는 점이다. 또한, 관찰기간 중 선택적으로 특정집단에서 중도 탈락이 발생한 경우 연구 결과가 왜곡될 우려가 있다(AHRQ, 2012).

2.1.1.2. 환자-대조군 연구

환자-대조군 연구(case-control study)는 특정 질병의 유무로 환자군과 대조군을 선정한 후, 과거 혹은 현재의 요인 인자 노출 상태를 조사하여 두 군 사이에서 비교하는 연구 방법이다. 환자군과 대조군 사이에 요인 노출의 정도 차이가 존재한다면, 그 요인이 질병 발생과 연관이 있다고 추론할 수 있다. 환자-대조군 연구와 코호트 연구의 가장 큰 차이점은 연구 대상자를 선별하는 방법에 있다. 코호트 연구가 질병이 없는 사람들 안에서 요인의 노출 유무로 집단을 구성하고 질병 발생까지 추적하는 반면, 환자-대조군 연구는 현재 질병의 유무로 집단을 구분한 뒤 과거 노출 여부를 조사한다. 이러한 방식 때문에 환자-대조군 연구는 후향적 연구라고도 불린다. 환자-대조군 연구에서 가장 중요하게 고려해야 할 사항은 연구 대상자 선택이다. 특히 대조군은 환자군이 발생한 동일한 관측시행 기반집단(source population)에서 요인의 노출 상태와 무관하게 선택되어야 한다. 그렇지 않을 경우 선택 바이어스(selection bias)가 발생할 수 있다.

환자-대조군 연구의 장점은 질병 위험군 전체를 연구하는 것이 아닌 위험 집단을 대표하는 샘플만 연구하기 때문에 효율적이고 결과를 빠르게 도출해 낼 수 있다는 점이다. 또한, 질병과 관련된 한 개 이상의 요인들을 조사할 수 있다.

그에 반해, 대조군이 전체 관측시행 기반집단이 아닌 그 안에서 선택된 일부 대상자들이므로 위험도(risk)나 위험차(rate differences)를 측정할 수 없고, 결과 해석이 상대적으로 어렵다. 또한, 위험요인에 대한 과거 노출 여부를 조사할 때 환자의 기억력에 의존하는 경우가 많은데, 대개 병을 가진 사람들이 그렇지 않은 사람에 비해 과거 노출력에 대해 더 잘 기억하므로 회상 바이어스(recall bias)가 발생할 수 있다(AHRQ, 2012).

2.1.1.3. 환자-코호트 연구

환자-코호트 연구(case-cohort study)는 환자-대조군 연구의 일종으로 관측시행 기반집단이 코호트인 연구 유형이다. 코호트 내의 모든 사람은 질병의 발생 유무나 코호트 내에 있었던 시간과 관계없이 동일한 확률로 대조군으로 선택될 수 있다.

환자-코호트 연구는 코호트가 관측시행 기반집단이기 때문에 추가 정보를 얻는데 있어서 환자-대조군 연구보다 효율적이며, 동일한 대조군의 사용을 통해 한 코호트 내에서 여러 질병에 관한 환자-대조군 연구의 수행을 용이하게 해준다.

반면에, 한 코호트 구성원이 환자군과 대조군 모두에 속할 수 있기 때문에 환자-

대조군 연구와 같은 수준의 통계적 정밀도(precision)를 얻기 위해서는 더 많은 수의 대조군을 필요로 한다(AHRQ, 2012).

2.1.1.4. 환자-교차설계 연구

환자-교차설계 연구(case-crossover study)는 환자-대조군 연구와 같이 질병을 가진 환자군을 선택하여 질병발생 전의 요인 노출에 대해 조사하지만 대조군을 따로 선정하지 않고 환자의 과거 노출력을 그들 자신의 대조군으로 삼는다. 각각의 환자들을 대상으로 서로 다른 기간 동안 노출 인자를 제외한 모든 변수들에 대한 별도의 측정이 이루어진다. 환자-교차설계 연구에서는 특정 질병을 가진 환자들 중 환자군 기간(case time period)과 대조군 기간(control time period)동안 노출 정도가 다른 환자만이 분석에 유효한 정보를 제공할 수 있다. 대조군 기간은 환자군 기간과 동일한 일수를 가져야하며, 개인별 짝짓기 환자-대조군 연구(individually matched case-control study)와 같이 분석된다.

환자-교차설계 연구의 장점은 대조군을 선택할 필요가 없고, 환자군과 대조군이 동일하므로 측정되지 않은 교란요인을 포함하여 시간이 지나도 변하지 않는 모든 교란효과를 통제할 수 있다는 점이다. 또한, 적은 비용으로 환자-대조군 연구의 추가연구로 진행될 수 있다.

반면에, 요인 노출력에 차이를 보이는 환자들만이 질병과의 상관관계 분석에 영향을 주기 때문에 비효율적일 수 있으며, 시간에 따라 변화하는 인자에 의해서는 교란효과가 발생할 수 있다. 비교효과 연구에서는 시간에 따라 변화하는 변수들 중 치료에 영향을 주거나 질병의 위험을 증가시키는 경우에 대해서 세심한 주의를 기울여야 한다(AHRQ, 2012).

2.1.1.5. 환자-시간-대조군 연구

환자-교차설계 연구의 경우 연구 집단의 노출에 대한 유병률은 시간에 따라 변하지 않는다고 가정하지만 실제로 이 가정은 쉽게 위배될 수 있다. 예를 들어, 노출이 치료법일 경우 새로운 치료법이 도입되었거나 기존의 치료법의 안정성에 문제가 제기되었을 경우 그 치료법의 유병률은 변화될 수 있다. 따라서, 이 문제를 해결하기 위해 환자-시간-대조군(case-time-control study) 연구방법이 제안되었다. 이는 환자-교차설계 연구에서 측정된 오즈비(odds ratio)를 대조군에서 측정된 상응하는 오즈비로 나누는 방식이다. 다시 말해, 환자-시간-대조군 연구는 환자-교차설계

연구에서 바이어스를 유발할 수 있는 치료 유병률의 달력시간 추세(calendar time trends)를 통제한다. 이를 위해 환자-대조군 연구에서와 같이 대조군이 선정되고, 환자군에서의 환자-교차설계 오즈비를 대조군에서의 환자-교차설계 오즈비로 나눈 값이 측정치로 이용된다.

환자-시간-대조군 연구는 환자-교차설계 연구가 가진 장점과 더불어 치료 유병률의 시간적 변화에 대한 가정에 영향을 받지 않는다는 추가 장점을 갖는다. 반면에, 환자-교차설계 연구에서 불필요했던 대조군이 다시 도입되며, 시간적 추세를 통제하는 과정에서 다른 교란효과를 유발할 수 있다(AHRQ, 2012).

2.1.2. 자료원

2.1.2.1. 자료원의 종류

관찰 연구를 이용한 비교효과 연구에 이용할 수 있는 자료원은 크게 두 가지로 나눌 수 있다. 첫째, 일차자료(primary data)는 연구자가 특정 연구를 목적으로 연구 참여 집단으로부터 직접 수집한 정보를 일컫는다. 자료는 직접(in-person) 인터뷰나 전화 인터뷰, 우편 조사와 같은 방법으로 얻을 수 있다. 관찰 연구의 경우 연구 가설을 측정할 수 있는 자료가 존재하지 않을 시 종종 새로운 자료의 수집을 필요로 한다. 일차자료 수집은 특정 연구 주제에 필요한 정보를 제공할 수 있다는 장점이 있는 반면, 시간이 오래 걸리고 비용이 많이 든다는 단점이 있다. 일차자료원의 예로는 등록자료(registry)를 들 수 있다. 예를 들어, 미국의 암등록자료인 SEER(-Surveillance, Epidemiology, and End Results)는 Medicare 자료와 연계하여 비교효과 연구에서 중요한 자료원으로 활용되고 있다.

둘째, 비교효과 연구에서 이용 가능한 자료로 이차자료(secondary data)를 들 수 있다. 이는 다른 목적으로 수집된 자료를 연구에 활용하는 것으로 전자 건강 기록(Electronic Health Record, EHR)이나 보험청구 자료가 이에 해당한다. 이차자료원의 경우 특정 연구를 목적으로 수집된 자료가 아니기 때문에 연구 목적에 맞는 질적 수준을 갖고 있는지에 대한 평가가 이루어져야한다. 또한, 각각의 자료원들이 제공하는 정보의 종류가 다를 수 있으며, 만약 환자가 여러 기관에서 의료 서비스를 받았을 경우 그 환자의 정보는 각각의 의료 기관이 갖고 있는 시스템에 입력되지만 별도의 통합된 정보를 제공하지 않는다(AHRQ, 2012).

2.1.2.2. DARTnet과 분산형 건강 데이터 네트워크

비교효과 연구의 시행을 위해서는 필요한 정보를 보다 체계적으로 제공할 수 있는 자료원이 필요하다. 일반적으로 관찰 비교효과 연구는 보험청구 자료와 같은 이차자료원을 많이 활용한다. 그러나 위의 설명과 같이 이차자료원은 특정 연구를 목적으로 수집된 자료가 아니기 때문에 여러 한계점을 지니고 있다. 특히, 연구에 필요한 질병의 중증도(severity)나 치료효과(clinical response)와 같은 임상정보가 부족하다. 이를 해결하기 위해 AHRQ는 기금을 조성하여 DARTnet(The Distributed Ambulatory Research in Therapeutics Network) 을 설립하였다. 이는 관찰 비교효과 연구 수행과 약물 처방과 의료 기기와 같은 의료 서비스의 효과에 대한 평가를 용이하게 하기 위해 설립된 전자 건강 데이터이다. DARTnet은 전자 건강 기록과 자가진단 자료(point-of-care)를 연계하여 연구에 필요한 다양한 정보를 제공한다. 2009년 이래, 100개 이상의 1차 진료 기관, 800명 이상의 임상의들과 100만명 이상의 환자를 포함하는 19개 기관들로부터 수집된 환자 단위 의료 데이터가 DARTnet을 통해 제공되고 있다(Libby 등, 2010).

또한, 비교효과 연구를 위해서는 여러 기관으로부터 수집된 전자 건강 자료가 필요하다. 전자 건강 자료에는 중앙통제형과 분산형 네트워크(distributed networks)를 이용한 방식이 있는데, 중앙통제형의 경우 자료의 보안문제나 소유권, 환자 정보의 침해와 같은 우려가 존재한다. 따라서 이에 대한 대안책으로 분산된 네트워크 방식으로 자료에 접근하는 시스템(Distributed Health Data Networks)이 구축되었다. 이는 자료 소유주가 보안 데이터와 데이터 사용에 대한 통제권을 그대로 유지하고, 연구자들로부터 필요한 자료에 대한 요청이 있을시 단일화된 형태로 자료를 반출하는 방식이다. 주로 이메일을 통해 제공된 동일한 컴퓨터 프로그램을 통해 자료 요청과 반출이 이루어진다. 이 방식의 자료원으로는 관리형 의료보험(HMO)의 Research Network Center for Education and Research on Therapeutics을 들 수 있다. 분산형 자료 체계는 중앙통제형 방식의 모든 기능을 수행하면서도 중앙통제형이 갖는 단점을 보완할 수 있다. 다시 말해, 자료에 대한 자료 소유기관의 통제가 유지되며, 이로 인해 보다 많은 자료 소유기관의 참여를 이끌어낼 수 있다. 또한, 환자 정보와 같이 보안이 필요한 자료에 대한 안전성이 강화되었으며, 여러 기관에서 수집된 모든 자료를 한꺼번에 보관할 중앙 시스템 역시 불필요해졌으며, 최신 정보를 저장하기 위해 수시로 업데이트를 해야 하는 번거로움도 줄일 수 있다. 따라서 비교효과 연구를 위한 대규모 분산형 건강 자료 네트워크의 단계적 도입이 제안되었다(Brown 등, 2010).

2.1.3. 관찰 연구 수행시 고려사항

비교효과 연구에서 관찰 연구의 경우 요인의 노출이나 치료가 연구자에 의해 정해지는 것이 아니라 일상적인 치료행위에서 자연스럽게 노출된다. 따라서 노출의 경위나 과정이 불분명하며, 무작위로 치료군이 나누어지는 임상시험에 비해 내적 타당도(internal validity)에서 쉽게 문제점이 야기된다. 내적 타당도란 관측시행 기반집단(source population)에 대한 추론(inference)의 타당성을 뜻하며, 연구 질문에 대해 얼마나 바이어스 없이 정확한 결과를 도출해 냈는지 여부로 평가된다. 내적 타당도를 약화시키는 요인으로는 크게 선택 바이어스(selection bias), 정보 바이어스(information bias), 그리고 교란 바이어스(confounding bias)가 있다(Rothman 등, 2008).

2.1.3.1. 선택 바이어스

연구 대상자 선택 과정이나 연구 참여율에 영향을 미치는 요인들에 의해 치료 효과 추정치가 왜곡되는 것을 의미한다. 대상자 선정과정에서의 바이어스는 주로 교란을 야기시키는데 이는 보정이 가능하지만, 대상자의 참여에 영향을 미치는 요인에 의한 바이어스는 보정이 불가능하다. 선택 바이어스를 야기시키는 가장 일반적 요소로는 연구에 참여한 집단과 연구에 참여하지 않은 집단을 포함한 연구 기반집단 간의 노출 요인과 질병과의 상관관계가 다르게 나타나는 경우이다. 특히 관찰 연구와 같은 비무작위 연구에서 선택 바이어스는 한쪽 중재에 배정된 참여자와 다른 군에 배정된 참여자 사이에 특성이 다를 경우 발생한다. 환자-대조군 연구에서 많이 발생하며 이 중에서도 병원 환자를 대조군으로 이용하는 경우에는 특히 주의해야 한다. 일반적으로 환자-대조군 연구에 비해 전향적 코호트 연구에서는 선택적 바이어스가 잘 발생하지 않지만, 근로자로 구성된 노출군과 일반인구로 구성된 비노출군을 비교할 경우 건강근로자 효과(healthy worker effect)라 불리는 선택 바이어스가 발생할 수 있다. 또한, 추적 기간 중에 노출군과 비노출군 사이에 차등적으로 탈락이 발생했을 경우에도 선택 바이어스 야기될 수 있다(안윤옥 등, 2005).

2.1.3.2. 정보 바이어스

노출 변수, 교란 인자 및 결과 변수에 대한 측정 오류가 치료 효과 추정치를 왜곡시키는 현상을 일컫는다. 측정 오류는 크게 비편향적(non-differential)과 편향적

(differential)으로 나타날 수 있다. 노출 변수나 결과 변수를 측정할 때 나타난 오류가 두 군 사이에 비슷하게 발생했을 경우를 비편향적이라 하며, 일반적으로 질병과 노출요인 사이의 관련성이 없는 방향으로 결과를 왜곡시킨다. 반면, 측정 오류가 두 군 중 한쪽으로 치우쳐 발생하는 경우를 편향적이라 하며, 이 경우 결과는 관련성이 없는 방향이나 관련성이 있는 방향으로 바이어스 된다. 정보 바이어스의 가장 대표적인 예로는 회상 바이어스를 들 수 있는데, 주로 설문 조사자의 편견이나 질병에 걸린 환자가 그렇지 않은 사람보다 자신의 과거 노출을 과장되게 기억하기 때문에 발생한다(안윤옥 등, 2005). 환자-대조군 연구에서는 질병 위험요인에 대한 과거 노출 여부를 설문지나 면접조사를 통해서 후향적으로 수집하기 때문에 회상 바이어스가 발생할 수 있다.

2.1.3.3. 교란 바이어스

노출변수와 결과변수의 관련성이 제 3의 변수에 의해 영향을 받아 왜곡되는 것을 말한다. 교란변수의 조건은 이 변수가 결과변수와 인과관계에 있어야 하며, 노출 변수와도 연관성이 있어야한다. 교란 바이어스는 연구설계 단계나 결과분석 단계에서 보정이 가능하나 측정되지 않은 교란요인에 의한 바이어스나 잔차 교란 바이어스(residual confounding)에 대해서는 추가적인 주의를 기울일 필요가 있다. 2개 이상의 치료법을 비교하는 비교효과 연구에서 적응증에 의한 교란(confounding by indication)은 가장 일반적인 문제점 중 하나이다(Sox and Goodman, 2012). 적응증에 의한 교란은 질병의 위중도(severity)나 환자의 연령에 따라 사용되는 치료법이 달라지고, 치료의 결과에도 영향을 주어 결과적으로 치료법과 치료결과의 연관성을 왜곡시키는 경우를 일컫는다. 예를 들어, 천식 환자의 사망률과 베타-애고니스트(beta-agonists)의 상관관계를 조사할 경우 천식의 심각도를 고려하지 않으면 중증의 천식환자에게 베타-애고니스트가 더 자주 처방되기 때문에 결과적으로 베타-애고니스트가 천식으로 인한 사망과 연관성이 있는 것처럼 보일 수 있다(AHRQ, 2012).

Note 교란 바이어스에 의한 문제점 해결 방안

교란 바이어스는 연구설계 단계와 자료분석 단계에서 보정이 가능하다. 연구설계 시 연구 대상자를 특정 집단으로 제한(restriction)하거나 잠재적 교란변수에 대해 매칭(matching)을 하면 교란변수의 영향을 통제할 수 있다. 자료분석 단계에서는 분석 대상자를 특정 집단으로 한정시키거나 교란변수로 의심되는 변수에 대해 층화(stratification)분석을 하여 교란영향을 보정할 수 있다. 또, 다양한 통계학적 방법이 교란요인의 영향을 통제하기 위해 사용되는데, 로지스틱 모형 등 다변수분석 (multivariable analysis)을 이용하여 교란변수의 영향을 수학적 모델 내에서 보정시키는 방법이 그 한 예이다 (안윤옥 등, 2005). 이 외에 성향점수(propensity score)와 도구변수(instrumental variables)들을 이용하여 교란영향을 보정할 수 있는데(Sox and Goodman, 2012), 이 두 방법에 대해서는 다음 장에서 자세히 설명될 것이다.

마지막으로, 민감도 분석(sensitivity analysis)을 사용하여 교란영향을 확인할 수 있다. 민감도 분석은 교란요인으로 의심되는 변수의 값을 변화시켰을 때 예측치(predicted outcome)의 변화를 측정하여 결과의 안정성을 확인하는 방법이다. 만약 결과값의 변화가 크다면, 그 변수는 결과에 중요한 영향을 미치는 요인이며 이에 대한 보정이 필요하다(Sox and Goodman, 2012).

민감도 분석은 측정되지 않은 교란요인(unmeasured confounder)의 영향을 평가하기 위해서도 사용될 수 있다. 측정되지 않은 교란요인이란 알려진 교란요인이지만 연구 데이터에 포함되지 않은 경우나 교란요인으로서의 관련성이 알려지지 않은 경우를 일컫는다. 비교효과 연구를 포함한 관찰 연구는 측정되지 않은 교란요인은 없다고 가정하나, 이 가정은 위배될 수 있으며, 관찰된 결과의 유효성에 영향을 미칠 수 있다 (AHRQ, 2012). 이 경우, 측정되지 않은 교란요인이 있다는 가설 하에 민감도 분석을 시행하여 연구 결과의 강건함(robustness)을 확인해 볼 수 있다. 관찰연구에서 교란을 야기시키는 요소로 비교군 선택이 있다. 예를 들어, 약물이나 호르몬과 같이 노출 기간이 중요한 변수를 연구할 경우 단순히 사용 여부만으로 대조군을 선택하는 것은 적합하지 않을 수 있다. 이 경우에는 한 개 이상의 비교군 (예. 비사용 집단, 가까운 과거에 사용한 집단, 먼 과거에 사용한 집단)을 포함시켜 각각의 결과 차이를 확인함으로써 연구 결과의 강건함을 확인해 볼 수 있다. 노출 변수나 결과 변수 혹은 공변량(covariate)의 정의에 의해서도 교란이 야기될 수 있다. 노출 변수의 경우 생물학적으로 질병 발생과 연관성이 있는 노출을 정의 내리기는 모호하다. 따라서 노출의 정도, 기간 및 빈도에 대한 정의를 변화시키면 노출과 결과변수 사이의 관련성 또한 변하게 된다. 이 경우, 노출의 정의를 변화시키면서 각각의 경우를 분석하여 결과의 안정성을 확인할 수 있다. 또, 임상적으로 적절한 결과 변수는 연구 자료에서 단일 진단 코드를 추출하거나 여러 개의 코드를 병합하여 정의내릴 수 있다. 이때 각각의 정의에 따라 분석을 시행하여 가장 적합한 결과 변수의 정의를 판단할 수 있다. 마지막으로 공변량이 잘못 분류되었을 경우, 교란효과를 유발할 수 있다. 예를 들어, 임상시험에서 사용되는 배정된 데로 분석하는 방법(intention-to-treat, ITT)은 관찰연구에서도 적용되고 있는데 이 때 민감도 분석을 통하여 ITT 집단과 실제 치료를 받은 집단을 비교하면 공변량의 분류가 결과에 미치는 영향을 확인할 수 있다(AHRQ, 2012).

2.1.4. 치료효과의 이질성

치료효과의 이질성(heterogeneity)이란 임상 결과로 측정된 치료 효과의 방향(direction)과 크기(magnitude)의 비무작위적 가변성(non-random variability)을 일컫는다. 관찰 연구를 통한 비교효과 연구에서 얻어진 치료 효과를 환자 개개인에게 적용할 시 이질성을 이해하는 것은 매우 중요하다. 예를 들어, 연령이 높은 환자층에서는 수술의 예후가 더 좋지 않거나 부작용이 발생할 확률이 높아질 수 있으므로 치료의 효과가 연령이 낮은 환자층보다 작게 나타날 수 있다. 혹은, 다수의 질환을 가진 환자의 경우 한 가지 이상의 치료 요법을 받고 있을 수 있으므로, 새로 적용된 치료의 효과가 다르게 나타날 수 있다. 이처럼 두 개 이상의 노출 요인이 합쳐져 질병이 야기될 경우, 노출 변수와 결과 변수 사이의 연관성이 노출 변수 이외의 제 3의 변수에 의해 다르게 나타나는데 이를 교호작용(interaction) 또는 효과변경작용(effect modification)이라고 한다. 효과변경작용이 존재할 경우, 효과변경인자(modifiers) 대한 적절한 고려가 이루어져야하며, 이 변수가 결과 변수에 대한 위험요인이 아닌 경우에도 마찬가지로 고려되어야 한다(AHRQ, 2012).

치료효과의 이질성은 연구 참여 집단의 특성이 다른 경우에도 발생할 수 있기 때문에 개별 연구들 간의 이질적인 결과에 대한 설명이 될 수 있다. 특히, 관찰 연구와 임상 연구 간에 나타나는 치료효과의 이질성에 대한 설명이 될 수 있는데, 이는 관찰연구의 경우 임상연구에 비해 다른 특성을 가진 환자들이 포함되기 때문이다.

치료효과의 이질성은 교란효과와는 달리 노출 변수와 결과 변수의 연관성을 왜곡시키지 않는다. 그러나 연구들 간의 비교를 위해서는 보정 효과를 통제할 필요가 있으며, 이때 가장 일반적으로 이용되는 방법이 하위그룹 분석(subgroup analysis)이다. 하위그룹 분석은 효과변경인자의 계층에 따라 그룹을 나누고, 각각의 그룹에 대한 치료 효과를 별도로 평가하는 방법이다. 교호작용의 통계적 검정(test for interaction)을 통해 효과변경인자와 치료법 사이에 교호작용이 존재한다고 판단될 경우, 효과변경인자의 카테고리(예. 남성과 여성) 별로 치료효과를 평가하게 된다(AHRQ, 2012).

교호작용의 통계적 검정을 수행할 시 몇 가지 주의할 점이 있다. 첫째, 일반적으로 하위그룹 간 치료효과의 이질성을 평가하기 위해서는 평균 치료 효과(average treatment effect)를 평가할 때 보다 더 큰 표본 크기(sample size)가 필요하다. 또한, 다수의 변수에 대해 교호작용을 테스트 할 경우, 제1종 오류가 증가하게 된다. 다시 말해, 유의수준(significance level)을 0.05로 놓았을 때 보고자하는 변수가 100개라면 이 중 5개의 변수는 실제로는 동일한 치료효과를 가짐에도 불구하고, 반

드시 통계적으로 유의한 치료효과의 이질성을 보이게 된다. 이러한 문제를 다중비교(multiple comparison)의 오류라고 하며, 이에 대한 보정이 필요하다. 마지막으로, 통계적인 교호작용 검정을 통해 얻어진 교호작용의 유무가 반드시 생물학적인 교호작용(biological interaction)의 유무를 의미하지는 않는다(AHRQ, 2012).

2.1.5. 비교효과 관찰연구의 사례

Sanoff 등(2012)은 일차 자료원과 이차 자료원이 연계된 환자 자료를 토대로 3기 대장암 환자를 대상으로 옥살리플라틴(oxaliplatin)이 첨가된 항암화학요법(oxaliplatin adjuvant chemotherapy)과 옥살리플라틴이 포함되지 않은 항암화학요법(non-oxaliplatin-containing adjuvant chemotherapy)의 효과를 비교분석했다. 기존의 임상시험에 따르면 5-플루오로라실(5-fluorouracil, 5-FU)에 옥살리플라틴을 첨가하면 3기 대장암 환자들의 생존률을 향상시킨다. 그러나 임상시험에 참가한 환자들은 일반 암 환자인구에 비해 비교적 연령이 낮고, 건강하며, 인종의 다양성을 반영하지 못한다는 한계점을 갖고 있다. 따라서 임상시험의 결과를 일반 암 환자집단에게 적용 가능한지에 대한 평가가 필요하다. 이 연구에서 사용된 자료원은 Medicare 청구자료와 연계된 SEER-Medicare, Medicaid와 Medicare 청구자료와 연계된 the New York State Cancer Registry(NYSCR-Medicaid, NYSCR-Medicare), the National Comprehensive Cancer Network(NCCN) Outcomes Database, 그리고 the Cancer Care Outcomes Research & Surveillance Consortium(CanCORS)이다. 연구 대상 집단은 5개의 자료원으로부터 선택된 75세 이하의 3기 대장암 환자 중 절제술을 받은 지 120일 이내에 항암화학요법을 받은 환자들이다. 선택된 환자들은 옥살리플라틴이 첨가된 항암화학요법을 받은 집단과 옥살리플라틴이 포함되지 않은 항암화학요법을 받은 집단으로 나누어졌다. 옥살리플라틴 치료집단의 경우 첫 번째 항암화학요법을 받고 30일 이내에 옥살리플라틴에 대한 청구나 처방 기록이 있는 환자들로 구성되었고, 그 외 나머지 환자들이 옥살리플라틴 비치료군에 포함되었다. 각각의 자료원 별로, 두 집단 사이의 생존률 차이가 측정되었고(즉, survival hazard ratios), 도출된 결과와 기존의 임상시험 결과의 비교분석도 시행되었다. 연구 결과, 기존 임상시험 결과와 마찬가지로 5-FU에 옥살리플라틴을 첨가한 경우 대장암 환자의 생존률이 증가하였다. 이와 같은 결과는 모든 자료원에서 동일하게 나타났는데, 특히 SEER-Medicare와 NYSCR-Medicare로부터 구성된 코호트의 경우 큰 표본크기를 갖기 때문에 통계적으로 유효한 결과를 보였다. 임상시험 결과들을 통합하여 얻어진 보정된 위험비

(adjusted hazard ratio, HR)에 따르면 옥살리플라틴의 치료군은 비치료군에 비해 대장암으로 사망할 확률이 약 20% 낮아지며(HR=0.80, 95% CI=0.70-0.92), SEER-Medicare에서 얻어진 자료에 따르면, 30%(HR=0.70, 95% CI=0.60-0.82), NYSCR-Medicare의 결과에 의하면 약 40%(HR=0.58, 95% CI=0.38-0.90) 정도 사망할 확률이 낮아짐을 알 수 있었다(Sanoff 등, 2012).

2.2. 성향 점수[1)]

무작위배정 비교임상시험은 무작위배정에 따라 치료가 결정되어 평가 대상인 치료 이외에 환자의 다른 특성들은 동일하고, 측정된 교란요인이 존재하지 않는다는 가정을 만족하기 때문에, 치료와 결과사이의 인과성을 설명할 수 있다. 하지만, 일반적인 관찰연구는 환자들의 기저특성이 다를 수 있고 측정되지 않은 교란요인도 있을 수 있기 때문에 치료방법들 간의 효과 비교가 어려울 수 있다.

성향점수는 연구대상 그룹들의 환자들 사이에 관찰된 공변량들의 균형을 맞춤으로써, 무작위배정 연구와 같은 상황을 만들어 치료효과를 비교할 수 있는 방법이다. 즉, 성향점수를 활용하면 결과에 대한 치료의 영향을 인과성으로 설명할 수 있다. 성향점수는 연령, 성별, 인종 등 측정된 다양한 변수를 합쳐 요약된 단일 변수로 만드는데 이 단일 변수를 성향점수라 한다. 성향점수를 이용하여 치료효과를 비교하기 위해서는 추정한 성향점수를 기준으로 치료군과 대조군을 배정하여 치료효과를 비교한다.

성향점수란 환자들의 관찰된 기저특성들이 공변량으로 주어졌을 경우 치료를 받을 조건부 확률로 정의되며, 공변량을 요약하는 균형점수(balancing score)이다. 성향점수는 흔히 로지스틱 회귀모형 또는 프라빗 회귀모형으로 추정할 수 있다. 추정된 성향점수는 제한(restriction), 매칭(matching), 층화(stratification), 공변량으로 보정하거나 가중치(weight)로 활용하는 방법으로 선택바이어스 문제를 해결한다.

성향점수 방법을 적용하여 바이어스가 없는 치료효과 추정치를 얻기 위해서는 두 가지 가정이 필요한데, 첫째로 치료효과에 영향을 줄 수 있는 모든 교란요인들에 대한 자료가 확보되어야 하고, 둘째로 반드시 치료를 받게 되는 환자와 반드시 치료

1) 장은진 등(2013), 측정된 교란요인을 고려한 성과분석 방법

를 받지 않게 되는 환자가 없다는 가정이 필요하다. 이 가정들을 만족할 경우 성향점수가 주어졌을 때 각 환자들은 무작위 배정 비교임상시험에서와 같이 치료를 배정받을 확률이 같게 된다.

성향점수 방법은 실제적인 치료효과를 평가하는 관찰연구에서 바이어스를 줄일 수 있는 사용하기 쉽고 매우 효과적인 분석방법이라고 할 수 있다. 특히 결과변수가 이분형인 경우 결과발생이 드물다면(예를 들어, 공변량 당 결과발생이 7개 보다 작은 경우), 성향점수 방법은 로지스틱 회귀모형에 비해 유용하다. 하지만, 성향점수 방법은 측정된 공변량만을 보정할 수 있다는 제한 점이 있다.

2.2.1. 방법

2.2.1.1. 성향점수 방법 적용 단계

성향점수 방법은 먼저 성향점수 모형에 포함될 변수와 추정방법을 결정하여 성향점수를 추정한 후 목표모집단과 치료군, 비치료군에 대한 성향점수의 분포의 겹치는 정도를 고려하여, 제한, 매칭, 층화, 공변량 보정, 가중치 방법 중 적절한 방법을 결정한 후 공변량의 균형을 확인해야 한다. 이때, 공변량들이 충분히 균형을 이루지 않은 경우 성향점수 추정 단계로 돌아가야 하며, 치료군과 비치료군이 균형을 이룬 경우 치료효과 및 치료효과의 분산을 추정하고 민감도 분석을 통해 치료효과를 비교한다. 단, 적용한 방법에 따라 치료효과가 민감한 경우 성향점수 추정단계로 다시 돌아가야 한다. 이와 같은 성향점수 방법의 적용단계를 요약하면 다음 그림과 같다.

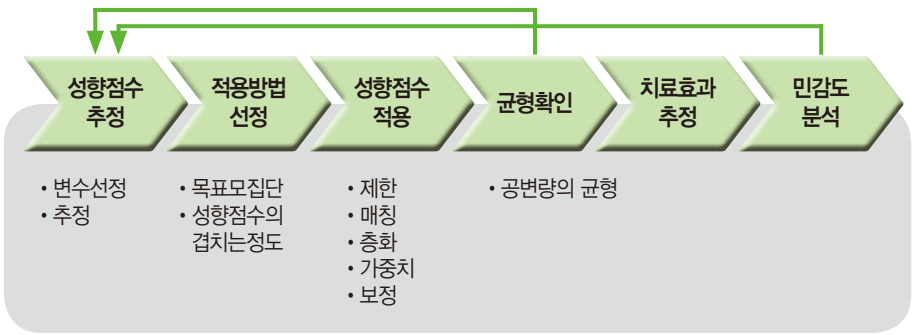

그림 2-1 **성향점수 방법의 적용 단계**
(자료원 : 장은진 등(2013), 측정된 교란요인을 고려한 성과분석 방법)

성향점수 추정모형에 포함될 공변량 선정시 교란요인은 반드시 포함해야 하며, 치료선택에 영향을 주지 않더라도 결과변수에 영향을 주는 변수는 분산을 감소시키는 효과가 있으므로 포함하는 것이 좋다. 그리고 치료배정에 따른 영향을 받을 수 있는 공변량은 성향점수 추정모형에 포함해서는 안 된다. 신뢰성있는 분석을 위해 결과분석 전에 성향점수 추정을 완료해야 하고, 어떤 성향점수 모형을 선택하느냐에 따라 결과에 영향을 받지 않도록 해야 한다(Rubin, 2007).

2.2.1.2. 성향점수 적용 방법

추정한 성향점수를 이용하여 제한, 매칭, 층화, 공변량 보정, 가중치 등의 방법을 적용하여 치료효과를 추정할 수 있다. 성향점수 적용 방법은 연구의 목적, 목표모집단, 치료군과 비치료군의 성향점수의 겹쳐지는 정도, 성향점수 적용 방법의 특징에 따라 결정할 수 있는데, 성향점수 적용방법의 특징을 요약하면 다음 표와 같다.

표 2-1 성향점수 적용방법의 특징

적용방법	특징
제한	– 치료효과 비교시 타당도 증가 – 평균치료효과 추정
매칭	– 성향점수 분포가 많이 겹쳐져 있을 경우 적용 가능 – 유사 무작위배정 역할을 함 – 가장 비교가능한 표본을 얻을 수 있음 – 치료효과 추정치의 바이어스가 작으나, 치료효과 추정치의 분산이 커짐 – 평균치료효과 추정
층화	– 성향점수 분포가 어느 정도 겹쳐져 있을 경우 적용 가능 – 전체 자료를 이용 – 치료효과 추정치의 바이어스는 커지나, 치료효과 추정치의 분산이 작아짐 – 평균치료효과, 치료군의 평균치료효과, 비치료군의 평균치료효과 추정
공변량 보정	– 결과변수가 드물게 발생하는 경우, 공변량들을 성향점수로 요약하여 보정할 수 있음 – 평균치료효과 추정
가중치	– 성향점수가 0 또는 1에 가까울 때 가중치가 커지는 문제점 있음 – 대상자 수가 상대적으로 작을 경우, IPTW 방법을 통해 치료군과 비치료군의 대상자수를 전체 대상자수로 증가시켜 평균치료효과를 추정할 수 있음 – 평균치료효과, 치료군의 평균치료효과, 비치료군의 평균치료효과 추정

(자료원 : 장은진 등(2013), 측정된 교란요인을 고려한 성과분석 방법론)

(1) 제한

제한방법은 성향점수 분포의 꼬리부분의 겹쳐지지 않은 영역, 즉 성향점수가 너무 높거나 낮아서 다른 치료군에 해당하는 대상자가 없는 경우를 찾아내어 제한하는 방법이다. 이처럼 성향점수의 차이가 큰 대상자를 포함하여 치료효과를 추정한다면 관찰된 자료의 범위를 관찰되지 않은 범위까지 확장하게 되므로, 만약 관찰되지 않은 영역의 변수들간의 관계가 변한다면 잘못된 결론을 내릴 수 있기 때문에 이를 제한하는 방법이다.

(2) 매칭

매칭방법은 치료군에 비해 비치료군의 환자수가 많고 치료군과 비치료군의 성향점수의 분포가 많이 겹쳐질 경우 사용하는 것이 적절하며, 매칭을 통해 치료를 받을 가능성이 낮은 환자(성향점수가 낮은 경우) 또는 항상 치료를 받게 되는 환자(성향점수가 높은 경우)는 제외된다. 때문에 매칭이 안 된 환자가 제외됨으로써 일반화(generalizability)가 어려워지고, 거의 모든 환자가 매칭이 안 되는 상황에서는 적절하지 못하다. 성향점수를 이용하여 매칭을 하는 경우 목표모집단은 치료군과 비치료군 전체이며, 매칭을 이용한 유사(pseudo) 무작위배정을 통해 평균치료효과를 추정할 수 있다.

(3) 층화

층화방법은 치료군과 비치료군의 성향점수의 분포가 어느 정도 이상 겹쳐 있을 경우, 전체 모집단에서의 효과성(effectiveness)를 파악하고자 할 때 사용할 수 있다. 이 때 모든 환자들의 자료를 사용하고 각 층의 치료효과를 통합할 때 층별 가중치를 조정함으로써 두 치료군의 평균치료효과 또는 치료군의 평균치료효과를 추정할 수 있다. Rosenbaum and Rubin(1984)의 연구결과 성향점수의 5분위수로 층을 나누었을 때, 성향점수 모형에 포함된 측정 교란요인으로 생기는 선형 치료효과 바이어스를 90% 정도 제거할 수 있는 것으로 나타나 일반적으로 5~10개의 하위분류가 가장 많이 사용되고 있다. 층화분석의 균형확인 방법은 매칭과 마찬가지로 각 공변량의 표준화의 차이를 비교하는 것이다.

(4) 공변량 보정

치료군과 비치료군의 성향점수의 분포가 거의 일부만 겹쳐지는 경우는 공변량 보정 방법을 선택할 수 있다. 결과의 발생이 드문 경우, 성향점수를 매칭하면 결과 발생이 더 줄어들고, 층화분석을 할 경우에도 층별로 결과의 발생이 작아서 치료효과

추정이 힘들 수 있다. 따라서 이런 경우는 공변량으로 성향점수를 포함시켜 보정하는 방법을 적용할 수 있다.

(5) 가중치

가중치 방법은 성향점수의 겹치는 정도와 상관없이 모든 환자들을 사용하는 방법이며, 목표모집단이 전체, 치료군, 비치료군 등 대상에 따라 가중치를 부여해 평균치료효과, 치료군의 평균치료효과 또는 비치료군의 평균치료효과를 추정할 수 있다.

많은 연구자들이 실제는 여러 방법을 적용하여 치료효과를 추정한다. 성향점수의 분포가 충분히 겹치는 경우에는 다른 방법을 적용하더라도 결과는 비슷해야 하며, 만일 결과가 다를 경우에는 자료를 다시 검토해야 한다. 추가적으로 성향점수 방법 적용시 성향점수모형에 포함되는 공변량을 변경하거나 다른 매칭방법을 적용하여 치료효과 추정치를 비교하는 민감도 분석이 필요하다.

2.2.2. 사례

2.2.2.1. 국내 사례

주천기 등(2010)이 발표한 「근시교정술의 장기간 안전성과 안정성 연구」에서는 LASIK수술과 표면절제술의 비교효과를 측정하기 위해 성향점수 방법을 활용하였다. 2002-2005년까지 서울 및 경기 지역의 6개 병원을 대상으로 후향적 의무기록 조사를 통해 근시교정술을 받은 5,109안에 대한 코호트를 구축하였다. 라식수술군과 표면절제술(라섹, PRK) 군 결정은 각막두께, 시력 등의 임상적인 특성에 따라 수술방법이 결정되므로 성향점수를 방법을 활용하였다.

성향점수를 사용하여 두 비교군간 균형 확보 후 유효성, 안전성, 안정성 비교과 표면절제술의 비교를 위해 연령, 술전 나안시력, 술전 교정시력, 안압, 현성굴절력, 현성굴절 구면도수와 난시도수, 각막두께 및 매칭 변수의 모든 교호작용 변수를 고려해 로지스틱 회귀분석으로 성향점수를 산출하여 그리드(greedy) 매칭방법을 사용해 1:1 매칭을 시행하였다. 그 결과 577안이 매칭이 되었고 수술의 유효성 지표로 술후 나안시력/술전 교정시력 비율을 보았고, 안정성 지표로 수술후 굴절력의 감소정도를 지표로 보았다. 그리고 안전성 지표로 불완전 각막판, 천공 절편형성, 안구건조증, 각막확장, 각막혼탁 등 분석하였다.

유효성 지표는 라식수술군(0.91)과 표면절제술군(0.86) 사이에 통계적으로 유의

한 차이가 없었으나 나안시력은 0.5미만으로 감소하고 각막혼탁의 누적발생률은 라식수술군보다 표면절제술에서 더 많이 나타났다.

Seung 등(2008)은 관상동맥 스텐트와 관상동맥우회술(Coronary Artery Bypass Grafting, CABG)의 효과를 비교하기 위해 성향점수를 활용하였다. 2000년 1월~2006년 6월 사이 한국에서 스텐트시술을 받은 1,102명과 CABG를 시행받은 1,138명에 대해 교란요인을 감소시키기 위해 성향점수의 매칭방법을 활용해 사망, 심근경색, 뇌졸중, 목표혈관의 재시술 등의 이상반응 결과를 비교했다. 환자 기본특성을 바탕으로 성향점수를 활용해 1:1매칭 코호트를 구축한 결과, 스텐트군과 CABG군 매칭은 각각 542명이었고, 금속 스텐트(Bare Metal Stent, BMS)군과 CABG군은 각각 207명, 약물방출 스텐트(Drug Eluting Stent, DES)군 vs CABG군은 각각 396명으로 매칭되었다. 각 그룹간 사망위험과 심근경색위험율은 통계적적으로 유의하지 않았으나 목표혈관 재시술 위험율은 스텐트군이 CABG군보다 유의하게 높게 나왔다(hazard ratio, 4.76; 95% CI, 2.80–8.11). 이러한 경향성은 약물방출 스텐트와 금속 스텐트 모두 비슷한 결과를 보였다.

Park 등(2010)은 ST가 상승된 급성 심근경색(ST segment Elevation Myocardial Infarction, STEMI) 환자에서 약물방출 스텐트와 금속스텐트의 치료효과를 비교하였다. 대상자는 2003~2006년까지 STEMI로 관상동맥 스텐트시술을 받은 684명이었다. 실제로는 539명이 약물방출 스텐트를 받았고 145명이 금속스텐트 시술을 받았다. 치료효과 결과로는 치료 실패와 관련된 심장마비, 재발, 재시술 등을 보았다. 기존 자료를 그대로 활용하여 다변수 보정을 한 후 콕스 비례 위험모델로 분석을 하였다.

하지만 이 자료는 관찰연구로 대상자의 기본 특성 등이 다르기 때문에 15개의 치료와 관련된 15개의 공변량으로 성향점수를 추정해 매칭방법을 사용하여 무작위 배정 임상시험 환경을 만들었다. 성향점수 매칭을 한 후 111명씩 약물방출 스텐트군과 금속 스텐트군으로 배정하였다. 그 결과 약물방출형 스텐트의 재시술율(감소는 통계적으로 유의했으나 비치명적 심근경색의 감소는 통계적으로 유의하지 않았다.

2.2.2.2. 국외 사례

(1) 매칭

Seeger 등(2007)은 1994년~1998년의 건강 보험 데이터베이스의 회원 중 고지혈증 환자를 대상으로 급성 심근경색(acute myocardial infarction) 환자에서 스타틴의 치료효과 대해 관찰연구 데이터로 성향점수의 매칭 방법을 적용하여 분석하였

다. 매칭으로 무작위배정을 하였고 성향점수로 확장된 변수조합을 확장 또는 감소시켰다.

스타틴의 노출은 모든 코호트 대상자의 연구기간 동안의 기록에서 수집했고 급성심근경색 결과는 청구자료의 청구코드를 기반으로 확인하였다. 성향점수 모델은 2번 수정했고 처음에는 심근경색과 관련 있는 38개의 위험요인 변수와 4개의 기간으로 설정했고, 두 번째는 52개의 변수, 6개의 기간으로 성향점수를 추정하였다.

건강보험 자료 대상자 중 504명이 스타틴에 노출되었고 8090명이 스타틴에 비노출된 대상자였다. 인구학적 특성만을 사용해 성향점수 매칭 방법을 통해 스타틴 치료군과 비치료군을 각 4,144으로 배정하고, 심근경색 관련 위험요인 38개의 변수와 4개의 기간을 성향점수화하여 스타틴 치료군과 스타틴 비치료군을 3,579명씩 배정하였다. 또한, 52개의 변수, 6개의 기간으로 성향점수를 추정해 치료군과 비치료군을 2,901명씩 배정하였다. 성향점수는 대부분 일치하였으나 변수 감소 및 몇 가지 변수에 대해서는 상당한 불균형을 보였다.

치료효과 추정 결과 성향점수에 기초한 매칭방법을 통해 교란요인이 완화되었고 성향점수 매칭의 스타틴 치료효과가 기존의 로바스타틴(lovastatin), 프라바스타틴(pravastatin), 심바스타틴(simvastatin) 등의 무작위 배정 비교임상시험과 유사한 결과를 보였다. 이러한 결과를 활용하기 위해서는 기존 임상시험 결과, 전문가 의견, 설문 조사, 차트 리뷰 또는 민감도 분석 등을 종합적으로 파악하여 결과를 제시하여야 한다.

(2) 가중치

Curtis 등(2007)은 비교효과분석을 위해 관찰연구에서 역확률 가중 추정량 활용방법에 대해 소개하였다. 역확률 가중치(Inverse Probability of Treatment Weighting, IPTW) 방법은 추정된 성향점수를 사용하여 목표모집단의 유사모집단을 만드는 방법이다. 무작위 비교임상시험에서는 3개의 β-차단제가 심부전 환자의 생존을 개선을 시키는 것으로 나타났다. 임상현장에서 의사들이 자주 직면하는 문제는 고혈압치료와 심부전 치료제로 기존 β-차단제를 사용한 경우 생존이 향상되었지만 비싼 신규 β-차단제로 바꿔야 하는 지가 고민거리였다. 심부전 환자에서 기존 약제와 신규 약제 두 치료방법 중 선택할 가능성은 균일하게 있기 때문에 역확률 가중 추정치 방법을 사용하기 위한 첫 번째 고려 상황은 충족된다. 두 번째로 하위그룹을 정의해야 한다. 만약 결과값으로 심부전으로 인한 입원 후 생존에 대해 관심이 있다면 입원전 β-차단제의 사용에 기초해 층화를 할 수 있다. 여기에서는 심부전으로 입원전 β-차단제를 사용하지 않은 환자를 분석대상으로 하여 하위그룹의 층

화없이 역확률 가중 추정치 방법을 사용하였다. 다음으로 성향점수를 추정하기 위해 모델을 만들고 나이, 인종, 민족, 동반질환(허혈성 심장질환, 고혈압, 심혈관질환, 고혈압, 폐질환, 당뇨병 등)등 치료결정과 결과에 관련 있는 모든 공변량을 포함하였다.

층화 분석은 정의한 하위 그룹이 다른 반응 특성을 보이는 경우에 적합하다. 예를 들어 관상 동맥 협착증의 치료를 위한 약물방출 스텐트 사용여부 결정 등에 적합하다. 금속 스텐트에 비해 약물방출 스텐트의 장기효과에 대한 근거가 부족함에도 불구하고 약물방출 스텐트는 미국의 관상 동맥 협회에서 우선순위를 차지하고 있다. 약물방출 스텐트의 빠른 사용을 감안할 때 일정 기간내 약물방출 스텐트를 받을 확률이 높을 것으로 예측된다. 이러한 것을 고려하여 모델에 기간을 정의하고 이 기간내 비교효과를 추정한다. 연구자는 성향점수의 분포를 확인하고 모델의 견고성을 확인하기 위해 공변량에 대한 민감도 분석을 반복해야 한다. 역확률 가중 추정치 방법은 관찰연구 데이터에 사용할 수 있는 강력한 도구이다. 하지만 주요 단점은 방법을 구현할 소프트웨어가 부족한 것이다.

(3) 층화

Segal 등(2007)은 약물군간 치료효과를 비교하기 위하여 경시적(longitudinal) 성향점수의 층화(subclassification)를 사용하여 시간에 따른 치료효과를 추정하였다. 당뇨병 환자를 대상으로 신약(exenatide)과 기존 당뇨병 치료제의 치료효과를 비교하였다.

2005년 6월~12월까지의 당뇨병환자 131,714명을 대상으로 성향점수를 4분위수로 층화하였다. 여기에서는 총 의료비용과 월간 의료비용을 추정, 비교하였다. 그 결과 평균 월간 의료비용의 차이는 $397(95% 신뢰 구간, $218–$1054)였으며, 입원의 상대 오즈(relative odds)는 1.02(95% CI, 0.33–1.98)였다. 그 결과 새로운 당뇨병 약물과 기존 치료약물의 치료효과 차이는 없었다. 시간이 지남에 따라 변화하는 치료방법을 포함하고, 관찰자료의 다면성을 감소시키 위해 경시적 성향점수의 층화방법을 사용하였다.

(4) 보정

Rassen 등(2010)은 성향점수를 이용하여 개인정보를 보호하면서 여러 자료원의 자료를 통합하여 보정하는 방법에 대해 소개하였다. 생물학적 안전과 관련된 대규모 연구들은 다양한 자료원을 통합한 결과를 요구하지만 교란변수를 조정하는데 필요한 개인정보는 공유할 수 없게 되어 있어 많은 어려움이 따른다. 이 문헌에서는

개인정보를 보호하면서 자료원을 통합해 모든 변수를 보정하는 방법으로 성향점수를 추정해 결과를 도출하고자 하였다.

개인 정보를 제외한 데이터는 중앙 데이터베이스로 전송되는 시스템을 구축하고 연구자는 전체 공변량을 공유하고 메타분석 및 성향점수 방법을 활용해 자료원을 통합하였다. 통합된 정보는 개인정보보호, 분석 무결성 및 유연성, 그리고 연구의 운영 및 통계 요구를 충족할 수 있는 지 등을 평가하였다.

보정된 성향점수와 개인의 공변량이 대부분 0~3%의 차이를 보인다면 이것은 유사하다고 간주할 수 있다. 이중 6개 이상의 노출 카테고리는 일반적인 노출 대표그룹과 반대그룹으로 구분하고 하위그룹 분석을 사전에 정의하였다. 그리고 시간 변화에 따른 노출과 교란변수 등은 성향점수를 재추정하는데 포함하였다. 그리고 각 자료원간 이질성에 대해 분석하였다. 성향점수 보정방법은 데이터베이스의 변수를 보정하고, 개인정보보호를 가능하게 하기 때문에 자료원 통합이 필요한 연구에 권장할 수 있다.

참고 계층모형

계층모형(hierarchical model)은 주어진 자료의 구조가 계층적 구조로 설명이 가능할 때 계층적 자료 구조의 특성을 반영하는 방법이다. 계층적 자료(hierarchical data)는 개체(subject)가 어떤 집단의 구조 내에 존재할 때 측정되는 자료이다. 예를 들면 가족내 구성원들의 건강상태를 측정한다면 그 측정값들은 계층적 자료가 된다. 여기서 가족은 집단이고 가족의 구성원은 개체가 되는 것이다. 계층적자료는 위의 예와 같이 2개의 층(level)만 가지는 경우도 있지만 3개 이상의 층을 가지는 경우도 있다. 예를 들면 어느 학교의 특정 반의 아이들의 시험성적의 경우, 학교(level 3) 내에 반(level 2)이 존재하고 그 반 밑에 학생(level 1)이 존재하는 경우이다. 같은 계층 내에 있는 개체들은 다른 계층의 개체들보다 유사함을 가진다. 이러한 유사함은 같은 계층 내에서 특정한 (환경적, 배경적, 경험적) 특징들을 공유하기 때문에 같은 계층으로부터의 관측값들은 독립성이 결여된다. 따라서 계층적 자료를 분석할 때는 계층적 자료의 특성을 반영할 수 있는 적절한 분석방법을 사용해야 한다.

계층모형 분석을 위한 모형들은 반응변수(response variable)의 형태에 따라서 다음과 같이 분류된다. 우선 반응변수가 연속형인 경우 자주 쓰이는 계층모형은 선형혼합모형(linear mixed model)이다. 선형혼합모형에서 랜덤변수는 각 층에 존재하는 계층의 변동을 설명해 주는 역할을 한다. 그리고 주어진 랜덤변수의 분산은 계층간의 이질성(heterogeneity)을 설명해 준다. 반응변수가 범주형 자료인 경우 주로 랜덤변수를 가지는 일반화 선형혼합모형(generalized linear mixed model)을 사용한다. 앞의 선형혼합모형과 같이 일반화 선형혼합모형에서도 각 층의 변동을 설명하는 랜덤변수를 가정하고, 그 변동이 각 층간의 다른 집단의 이질성을 설명해 준다.

위에 제시한 계층모형의 추정치 및 예측값을 계산하기 위해서 많은 통계패키지들이 이용되고 있다. 대표적인 것이 SAS에서 Proc mixed, Proc glimmix, 그리고 Proc nlmixed이다. Proc mixed는 선형혼합모형의 분석을 위해서 사용되고, Proc glimmix와 Proc nlmixed는 일반화 선형혼합모형을 위해서 이용된다. 이러한 방법은 전통적 통계방법에 기초한 방법이고, 베이지안(Bayesian) 방법에 기초한 패키지는 WinBUGS가 있다.

Normand 등(2010)은 고관절 치환술에 들어가는 의료기기의 효과에 대한 연구에서 시판전(premarket) 연구들의 결과와 시판후(postmarket) 연구들의 결과를 통합하여 결론을 도출하였다. 이 연구에서 결과변수가 고관절 치환술 후 활동 및 기능의 향상 정도를 나타내는 점수이므로 계층적 일반화 선형혼합모형을 사용하여, 층내에서의 효과를 랜덤효과로 고려하였다.

계층모형은 비교효과연구에서 계층적 자료 구조를 가지는 경우 폭넓게 사용할 수 있는데, 예를 들어 체계적 문헌고찰 결과 여러가지 연구설계의 연구들의 결과를 통합할 필요가 있을 경우에도 연구설계를 층으로 고려하여 계층모형을 적용할 수 있으며, 연구자료가 여러 지역, 여러 기관들에서 수집되어 이들 층내 유사성을 고려할 필요가 있는 경우에도 계층모형을 적용할 수 있다.

참고 경시적 자료 분석

경시적연구(longitudinal study)는 각 개체(subject)가 일정기간의 시간동안 반복적으로 측정되는 연구이다. 따라서 경시적연구는 각 개체에게서 한 개의 결과치만 가지는 횡단면연구(cross-sectional study)와는 구분이 되는 것이다. 경시적연구에서 만들어진 반복측정자료를 경시적자료(longitudinal data)라고 하고, 이러한 경시적자료의 특성은 다음과 같다. 첫째, 경시적연구에서는 반복측정으로 자료의 수가 증가하고, 그 결과 검정력이 증가한다. 둘째, 개체들의 비교 또는 개체가 속한 그룹들 간의 비교를 위한 추론을 강조한다. 셋째, 경시적자료는 개체들의 반복측정으로 인한 상관관계가 존재하기에 이 상관관계를 적절하게 모형화하는 것을 종종 요구한다. 넷째, 측정으로 인하여 개체간에 그리고 시간에 걸친 변동을 가지게 된다. 이러한 경시적자료의 변동(variation)은 3가지로 구별되는데, 1) 개체들간의 이질성(heterogeneity between subjects), 2) 시간적 상관관계(serial correlation), 3) 측정오차(measurement error)이다. 경시적자료분석은 이러한 변동을 잘 설명해 줄 수 있다. 마지막으로 경시적연구에서는 개체들의 시간에 따른 효과(effect)의 변화를 추정할 수 있다. 위의 특성들은 횡단면연구에서는 가질 수 없는 경시적자료만이 가지는 장점이다 (Diggle 등, 2002).

위에서 설명했던 바와 같이 경시적자료는 횡단면연구와 다르게 반복측정으로 인한 상관관계를 가지게 된다. 만약 이러한 상관관계를 무시하고 횡단면연구에서 하는 방법으로 회귀계수들을 추정하게 되면, 추정된 회귀계수는 심각한 바이어스를 가질 수 있다. 그리고 표준오차(standard error)값이 큰 비효율적인 추정치를 가질 수도 있다(Diggle 등, 2002). 따라서 이러한 반복측정된 결과치들의 상관관계 올바르게 처리하면서 회귀계수를 정확하게 추정하는 모형들이 많은 통계학자들에 의해서 제안되어 왔다.

경시적자료분석을 위한 모형들은 반응변수의 형태에 따라서 다음과 같이 분류된다. 우선 연속형 반응변수를 위한 대표적인 모형으로는 분산분석(analysis of variance, ANOVA) 형태의 분석과 선형혼합모형이 있다(Hedeker and Gibbons, 2006). 분산분석 형태의 분석은 오차항이 정규분포를 따르고 등분산을 가정한다. 그리고 그룹간의 평균비교와 추정에 관심을 가진다. 하지만 시간의 효과를 범주형으로 간주하고 연속형 독립변수를 사용하는데는 한계가 있다. 이러한 단점을 극복하기 위해서 선형혼합모형이 제안되었다. 선형혼합모형은 시간변동에 따른 개별 개체의 변동을 모형화한다. 그리고 분산분석 형태의 분석보다는 분산의 형태를 좀더 일반적인 형태로 가정하는 모형을 가진다. 예를 들면 분산분석 형태의 분석에서는 가질 수 없는 공분산행렬의 구조인 자기상관(autoregressive), 혼합대칭(compound symmetry) 등을 가질 수 있다. 따라서 요즘은 주로 선형혼합모형을 주로 많이 사용하고 있다(Diggle 등, 2002). 이러한 모형의 추정을 위해서 주로 SAS에서 Proc glm과 Proc mixed가 사용된다.

범주형 반응변수를 분석하기 위한 모형은 개체특성적효과(subject-specific effect)를 위한 조건부모형(conditional model)과 모집단의 평균적인 효과(population-averaged effects)를 위한 주변모형(marginal model)으로 구분된다(Daniels and Hogan, 2008). 조건부모형(conditioanl model)은 반복측정에 의한 상관관계를 설명하기 위한 랜덤변수 또는 그전 시간에 관찰된 반응변수의 값을 이용한 마코프구조를 이용하고, 반응변수의 평균과 설명변수의 관계를 간접적으로 설명하는 모형이다. 대표적인 모형이 일반화 랜덤효

과 모형(generalized random effects model)과 전이모형(transitional model)이 있고, 이 두 모형들은 모두 우도함수(likelihood function)를 기초로 한 모형이다. 주변모형의 경우는 평균과 분산, 주변밀도함수(marginal density function), 그리고 작업상관행렬(working correlation matrix)로 주변모형에 있는 설명변수의 회귀계수를 일반화 추정방정식(generalized estimating equation, GEE)을 이용하여 추정한다(Agresti, 2002). GEE에 의한 모수추정은 계산이 간단하고, 일반화 선형모형(generalized linear model, GLM)하에서 해석을 하기에 많은 연구자들이 쉽게 사용할 수 있다. 하지만 우도원리를 기초로 한 방법이 아니므로 우리가 흔히 쓰는 우도비검정(likelihood ratio test)과 벌점에 기초한 모형선택기준들(penalized model selection criterion)인 AIC(Akaike information criterion) 또는 BIC(Bayesian information criterion) 등을 사용할 수 없다. 이러한 모형들의 추정을 위해서 SAS에서는 Proc glimmix, Proc nlmixed, 그리고 Proc genmod가 사용된다.

실제 임상현장을 반영하는 비교효과연구에서는 동일 대상자에게 반복적인 조사나 측정을 통해 자료를 수집하는 경우가 많이 발생한다. 이때 반복측정된 자료를 이용하여 효과를 추정하고자 할 경우, 경시적 자료의 특성을 반영한 분석방법을 적용하여 바이어스가 작은 효과를 추정하는 것은 매우 중요하다고 할 수 있다. 따라서 연구목적과 연구자료에 따라 적합한 분석방법을 적용하여야 한다.

참고문헌

안윤옥, 유근영, 박병주 외. 역학의 원리와 응용. 서울대학교출판부. 2005.

장은진, 안정훈, 정선영, 황진섭, 이자연, 심정임. NECA 측정된 교란요인을 고려한 성과 분석 매뉴얼. 한국보건의료연구원. 2013.

주천기, 차흥원, 현준영, 김미금, 김태임, 김진형, 김재용, 정소향, 나경선, 변용수, 김진국, 조은영, 김응권, 김재훈, 이자영, 김세경, 최지은, 장은진, 정선영, 이은주, 이나래. 근시교정술의 장기간 안전성과 안정성. 한국보건의료연구원. 2010.

Agresti A. Categorical Data Analysis. 2nd edition. Wiley, 2002.

Ahn J. Beyond Single Equation Regression Analysis: Path Analysis and Multi–Stage Regression Analysis. AM J PHARM EDUC 2002; 66(Spring):37–42.

AHRQ. 2012. Developing a Protocol for Observational Comparative Effectiveness Research (OCER): A User's Guide.

Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity–score matched samples. Stat Med. 2009 Nov;28(25):3083–3107.

Brown J, Brown JS, Holmes JH, et al. Distributed health data network: a practical and preferred approach to multi–institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010; 48(suppl 1):S45–S51.

Brookhart MA, Rassen JA, Wang PS, et al. Evaluating the Validity of an Instrumental Variable Study of Neuroleptics: Can Between–Physician Differences in Prescribing Patterns Be Used to Estimate Treatment Effects? Medical Care. Oct 2007; 45(10 Supp 2): S116–S122.

Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiology and Drug Safety 2010; 19: 537–554.

Concato J, Lawler E, Lew RA, et al. Observational Methods in Comparative Effectiveness Research. The American Journal of Medicine. 2010;123:e16–e23.

Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. Am J Med. 2010 Dec;123(12 Suppl 1):e16–23.

Concato J, Peduzzi P, Huang GD, et al. Comparative Effectiveness Research: What Kind of Studies Do we Need? Journal of Investigative Medicine. 2010;58:764–

769.
Curtis LH, Hammill BG, Eisenstein EL, et al. Using inverse probabilityweighted estimators in comparative effectiveness analyses with observational databases. Med Care. 2007;45(Suppl 2):S103–S107.
Daniels MJ, Hogan J. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall/CRC, 2008
Diggle PJ, Heagerty P, Liang K, Zeger SL. Analysis of Longitudinal Data. Oxford Press, 2002.
Sox HC and Steven N. Goodman. The Methods of Comparative Effectiveness Research. the Annual Review of Public Health. 2012;33:425–445.
Hedeker D, Gibbons RD. Longitudinal Data Analysis. Wiley, 2006
Libby AM, Pace W, Anderson HO, et al. Comparative effectiveness research in DARTNet primary care practices: point of care data collection on hypoglycemia and over–the–counter and herbal use among patients diagnosed with diabetes. Med Care. 2010;48(suppl 1):S39–S44.
MacPherson H. Pragmatic clinical trials. Complementary Therapies in Medicine. 2004; 12: 136–140.
MacPherson H, White A, Cummings M, Jobst K, Rose K, Niemtzow R. Standards for reporting interventions in controlled trials of acupuncture – the STRICTA recommendations. Comp Ther Med 2001; 9(4): 246–9.
McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? : Analysis using instrumental variables. JAMA. 1994; 272(11): 859–866.
Newhouse JP and McClellan M. Econometrics in outcomes research: The use of instrumental variables. Annu. Rev. Public Health. 1998; 19: 17–34.
Normand S, Marinac–Dabic D, Sedrakyan A. Rethinking analytical strategies for surveillance of medical devices: the case of hip arthroplasty. Med Care. 2010;48(suppl 1):S58–S67.
Park KW, Kang SH, Chung WY, Lee HY, Park JS, Kang HJ, Cho YS, Youn TJ, Koo BK, Chae IH, Choi DJ, Hahn S, Park BJ, Kim HS. 'Real world' comparison of drug–eluting stents vs bare metal stents in the treatment of unselected patients with acute ST–segment elevation myocardial infarction. Circ J. 2010 Jun;74(6):1111–20. Epub 2010 Apr 20.
Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73:1–11.

Rassen JA, Solomon DH, Curtis JR, et al. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. Med Care. 2010;48(suppl 1):S83-S89.

Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 1984 Sep;79(387):516-24.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983 Apr;70(1):41-55.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology: Third Edition. Lippincott Williams & Wilkins. 2008.

Sanoff HK, Carpenter WR, Martin CF, et al. Comparative Effectiveness of Oxaliplatin vs Non-oxaliplatin-containing Adjuvant Chemotherapy for StageIII Colon Cancer. JNCL. 2012;104:1-17.

Schafer JL, Kang J. Average causal effects from nonrandomized studies: A practical guide and simulated case study. Psychol Methods. 2008 Dec;13(4):279-313.

Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates: an example and lesson. Med Care. 2007;45(Suppl 2):S143-S148.

Segal JB, Griswold M, Achy-Brou A, et al. Using propensity scores subclassification to estimate effects of longitudinal treatments: an example using a new diabetes medication. Med Care. 2007;45(Suppl 2): S149-S157.

Seung KB, Park DW, Kim YH, Lee SW, Lee CW, Hong MK, Park SW, Yun SC, Gwon HC, Jeong MH, Jang Y, Kim HS, Kim PJ, Seong IW, Park HS, Ahn T, Chae IH, Tahk SJ, Chung WS, Park SJ. Stents versus coronary-artery bypass grafting for left main coronary artery disease. N Engl J Med. 2008 Apr 24;358(17):1781-92. Epub 2008 Mar 31.

Stuart EA. Matching methods for causal inference:A review and a look forward. 2009

Suissa S. The case-time-control design. Epidemiology. 1995;6(3):248-53.

Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models:the case of the omitted confounder. Pharmacoepidemiol Drug Saf. 2005 Apr;14(4):227-38.

PART 3

체계적 문헌고찰

3.1. 개요

체계적 문헌고찰(systematic reviews)은 특정 연구 질문에 대해 연구 시점에서 사용 가능한 최선의 연구 결과를 종합하는 연구 방법이다. 체계적 문헌고찰은 체계적이고 포괄적인 문헌검색과 사전에 정해진 포함/배제 기준에 따른 문헌선택, 선정된 문헌에 대한 비뚤림 위험 평가 등의 엄격하고 객관적인 연구 과정을 거친다(김수영 등, 2011).

체계적 문헌고찰은 지난 25년간 가장 중요한 방법론 중 하나로 발전하여 왔으며, 이러한 체계적 문헌고찰은 임상지침, 임상평가 등 임상현장에서의 의사결정과 보건정책 결정과정, 그리고 공중보건사업 등을 만들어가는 과정에서 일차연구 결과를 근거로써 사용하고 있다. 예를 들어 미국의학협회(Institute Of Medicine, IOM)에서는 최근 임상지침의 표준에 대한 보고서에서 권고사항을 결정할 때 가장 처음으로 체계적 문헌고찰을 수행하라고 하였다(Sox and Goodman, 2012).

비교효과연구에서의 체계적 문헌고찰은 비교효과 체계적 문헌고찰(Comparative Effectiveness Review, CER)로 표현된다. AHRQ에서는 2012년 4월 비교효과 체계적 문헌고찰에 대한 방법론을 정리하여 비교효과 체계적 문헌고찰의 특징에 대한 보고서 초안(draft)을 발간하였다. 이 보고서에 따르면 비교효과 체계적 문헌고찰은 어떤 단일 치료(single therapy)가 안전하고 효과적인지에 대한 좁은 범위의 연구 질문에 대답하는 것이 아니라, 의사와 환자가 만나는 임상의료 현장에서의 비교 대안들 중 어떤 것이 얼마나 상대적으로 이득이 있고 해로운지를 연구하는 것이다. 따라서 비교효과고찰에서는 실제 임상상황을 반영하는 것이 매우 중요하며, 이를 위해서는 임상상황에 대한 정확한 이해가 필요하다. 임상 현장에서의 비교 대안의 이득과 위해에 대해 결론을 내리기 위해서는 시험군과 대조군의 직접비교 효과시험(head-to-head effectiveness trial)의 연구 디자인이 가장 좋으나, 일반적으로 이러한 연구 디자인으로 수행된 연구가 충분한 경우는 거의 없다. 이럴 때는 위약-대조 시험(placebo-controlled trial)이나 관찰연구(observational study)가 유용하다. 특히 관찰연구는 비교 대안들의 위해(harm), 순응도 및 지속성 등을 볼 때 유용하다. 또한 직접비교 시험이 없다면 간접비교(indirect comparison)가 유용한지 여부를 검토하는 것이 좋다. 간접비교에 대해서는 4.3절을 참고하기 바란다.

실제 임상현장에서의 적용을 위해 비교효과 체계적 문헌고찰에서 도출한 근거의 적용 가능성을 주의 깊게 살펴보는 것은 중요한 일이다. 유용한 비교효과 체계적 문

헌고찰은 여러 대안들 사이에서 이득(benefit)과 위해를 비교한 것으로, 각각의 대안은 대상자와 환경에 따라 다양해질 수 있다. 여기에서 위해에 대한 근거는 일반적으로 무작위배정비교임상시험에서 알아내기가 어려운데 이러한 점에서 관찰연구는 임상시험에서 관찰되는 결과들과는 달리 실제 임상 현장에 적절한 세팅과 환자들 사이에서 이득과 위해가 어떻게 나타나는지, 임상시험과는 다른 관점에서 체크할 수 있다는 장점이 있다.

3.2. 방법

일반적으로 체계적 문헌고찰의 연구과정은 아래 그림과 같은 흐름을 따른다.

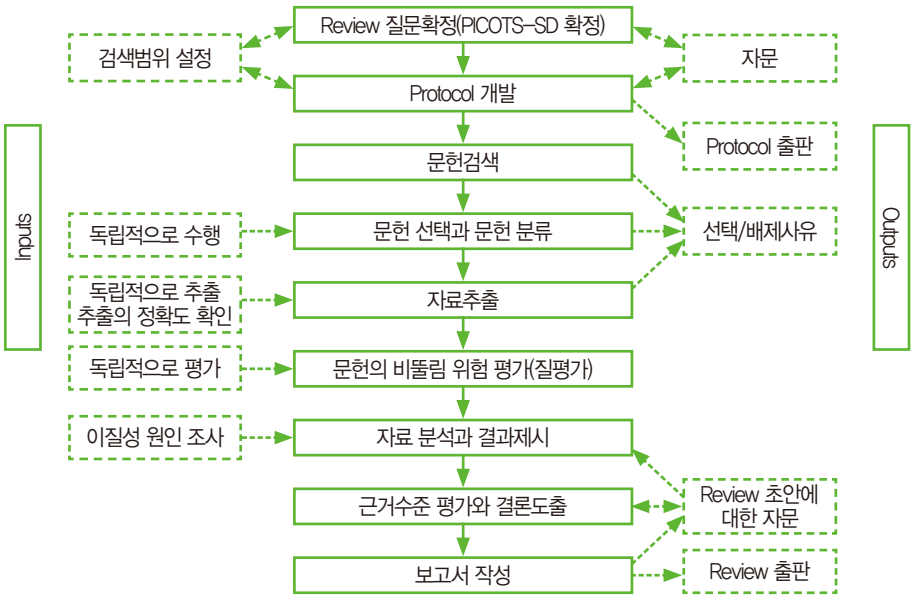

그림 3-1 체계적 문헌고찰 연구과정 흐름도

(자료원 : 김수영 등(2011), NECA 체계적 문헌고찰 매뉴얼)

체계적 문헌고찰의 각 과정에서의 구체적인 방법론은 「NECA 체계적 문헌고찰 매뉴얼(김수영 등, 2011)」을 참고하도록 한다. 여기서는 주로 비교효과연구에서의 체계적 문헌고찰, 즉 비교효과 체계적 문헌고찰에 중요한 사항을 다루도록 한다.

3.2.1. 연구기획과 프로토콜 작성

AHRQ의 비교효과 체계적 문헌고찰 보고서 초안(2012)에 따르면 비교효과 체계적 문헌고찰을 잘 수행하기 위해서는 다음 사항들을 고려해야 한다.

- 실제 임상적이고 환자 중심의 관점으로 근거에 접근한다.
- 어떤 의료 서비스에 있어서 '근거'를 기반으로 하고, 또한 임상적인 논리를 잘 찾도록 한다.
- 다양한 형태의 근거와, 효과에 관한 근거의 질과, 적용수준에 따라서 근거 수준을 계층화한다.
- 정책 결정자들이 다양한 치료법과 진단의 전략들 사이에서 공정하게 비교할 수 있도록 현재의 치료법과 비교 대안들의 이득과 위해에 대한 정보를 일관된 방식으로 제공한다.
- 비교효과 체계적 문헌고찰은 언제나 실제 임상현장에 적용하기 위한 목적을 가지고 있고 실제 임상현장을 반영하도록 하고 있다. 실제 임상현장에 대한 적절한 근거가 없거나 부적절할 때는 연구자들 간 의견의 차이를 줄이기 위해 어떠한 방법을 취해야하는지 미리 정의하도록 한다.

비교효과 체계적 문헌고찰에서 다른 체계적 문헌고찰에 비해 강조하는 점은 실제 임상 현장에서의 효과를 중요하게 생각하고 환자의 관점에 초점을 맞추어야한다는 점이다. 이를 위해서는 밝히고자 하는 연구의 질문을 명확하게 하기 위해 분석의 모형을 세우는 것이 도움이 된다. 미국 AHRQ의 EPC(Evidence-based Practice Centers) 프로그램에서는 비교효과 체계적 문헌고찰 수행 시 분석의 틀(analytic framework) 혹은 개념적 틀(conceptual framework)을 많이 사용한다. 이러한 틀을 사용하게 되면 이를 통해 대리지표 혹은 중간지표와 결과지표(health outcomes)의 관계를 명확히 할 수 있고, 또한 연구하고자 하는 중재에서 오는 이득에 대한 가정을 명확하게 할 수 있다. 예를 들면 대리/중간지표로는 콜레스테롤 수치가 될 수 있고 결과지표로는 심근 경색이나 뇌졸중과 같은 질병의 이환이 될 수 있다. 중재에서 오는 이득으로는 삶의 질, 이환율, 사망률 등 장기간 효과에 대한 평가로서 이루어질 수 있다.

예를 들어 Talati 등(2011)은 항간질약(antiepileptic medication)의 유효성, 안전성, 내성(tolerability)을 확인하고 예전 약과 새로운 약을 비교 평가하고자 하는 연구이다. 항간질약의 효과성과 안전성을 평가하기 위해 〈그림 3-2〉와 같은 분석의

틀을 제시하였고 이 분석의 틀에 따라 다음의 핵심 질문 네 가지를 도출하였다.

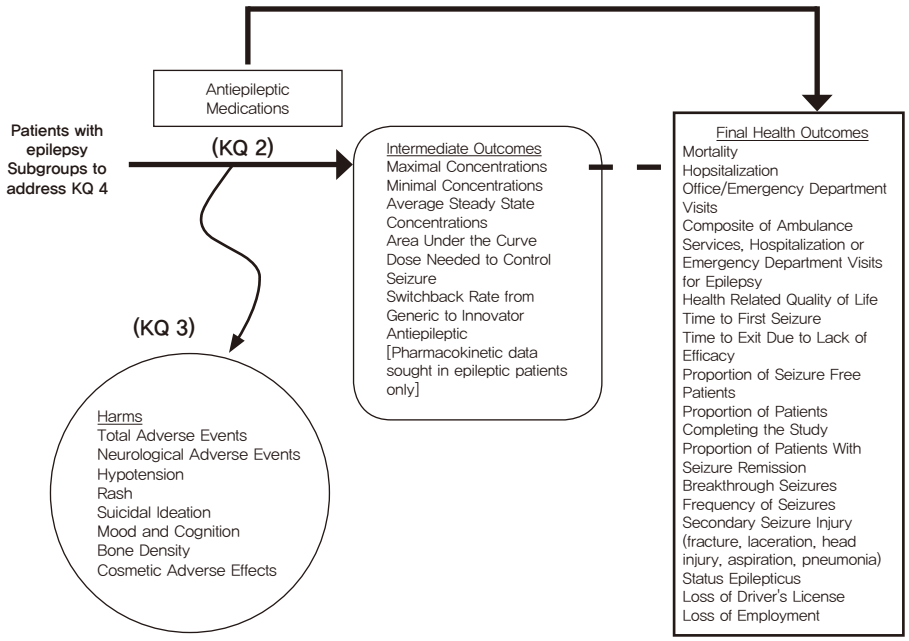

그림 3-2 간질 환자에게서 항간질약의 효과를 평가하는 분석의 틀

(자료원 : Talati 등(2011), Effectiveness and Safety of Antiepileptic Medications in Patients With Epilepsy)

핵심질문 1 : 간질이 있는 환자에게 어떤 약이 건강 결과지표에 대해 효과적/효능적인가? 여기서는 효과와 효능을 같이 보기로 하였고, 건강 결과지표로는 사망률, 입원, 외래나 응급실 방문, 건강 관련 삶의 질, 경련 등으로 정하였다.

핵심질문 2 : 간질이 있는 환자에게 어떤 약이 대리지표에 대해 효과적/효능적인가? 대리지표로는 약동학(pharmacokinetics), 경련을 조절할 때 필요한 약물의 용량, 약의 변경율(switchback rate)을 보았다.

핵심질문 3 : 간질이 있는 환자에게 어떤 약이 심각한 이상반응이 있는가? 심각한 이상반응으로는 신경학적 부작용, 고혈압, 발진, 자살관념(suicidal ideation), 인지기능 변화, 골밀도 등으로 하였다.

핵심질문 4 : 간질이 있는 환자들을 그룹별로 분류하였을 때 어떤 약이 어떤 이득과 위해가 있는가? 그룹은 경련의 원인과 형태, 성별, 인종, 연령 등으로 분류하였고 항경련제 또한 분류하였다.

이렇게 분석의 틀을 만들어 보는 것은 이 연구에서 보고자 하는 것이 무엇인지, 그리고 실제 연구 가능한 핵심 질문은 무엇인지를 도출하는데 도움을 준다. 따라서 연구 기획 단계에서 분석의 틀을 만들어보는 것이 좋다. 분석의 틀을 만들 때는 임상적, 정책적 판단이 필요하므로 해당 질환의 환자, 임상 전문가, 정책 결정자 등 이해 관계자들이 관여하도록 한다.

이렇게 핵심 질문이 정해지면 이에 따라 선정/제외기준을 설정해야한다. 비교효과 체계적 문헌고찰에서는 관찰연구와 같은 다양한 연구 디자인이 포함될 수 있기 때문에 항상 양적 합성인 메타분석이 가능하지 않을 수 있다. 이러한 부분을 잘 고려하여 선정/제외기준을 잘 정해야하며 이에 따라 프로토콜을 작성하도록 한다.

3.2.2. 문헌선택 및 분류

비교효과 체계적 문헌고찰의 문헌 선택 및 분류에 대한 이슈는 일반적인 체계적 문헌고찰과 크게 다르지 않다. 다만 비교효과 체계적 문헌고찰의 특성상 관찰연구를 포함하여 검색하는 경우가 많은데 이 경우, RCT만을 포함하는 체계적 문헌고찰보다 넓은 검색 범위로 인해 보다 많은 문헌이 선정될 것이다.

비교효과 체계적 문헌고찰에서는 관찰연구를 포함하는 경우가 많은데 AHRQ에서는 관찰연구를 포함하고자 결정하는데 도움이 되는 틀을 제안하고 있다(그림 3-3). 이 틀에 의하면 두 가지 질문을 통해 관찰연구를 포함할지 결정하도록 하고 있다.(Norris 등, 2011)

첫째, 연구 질문에 대해 RCT로부터 나온 근거가 빈틈(gap)이 있는가?

둘째, 관찰연구가 핵심질문을 설명하는데 타당하고 유용한 정보를 제공할 것인가?

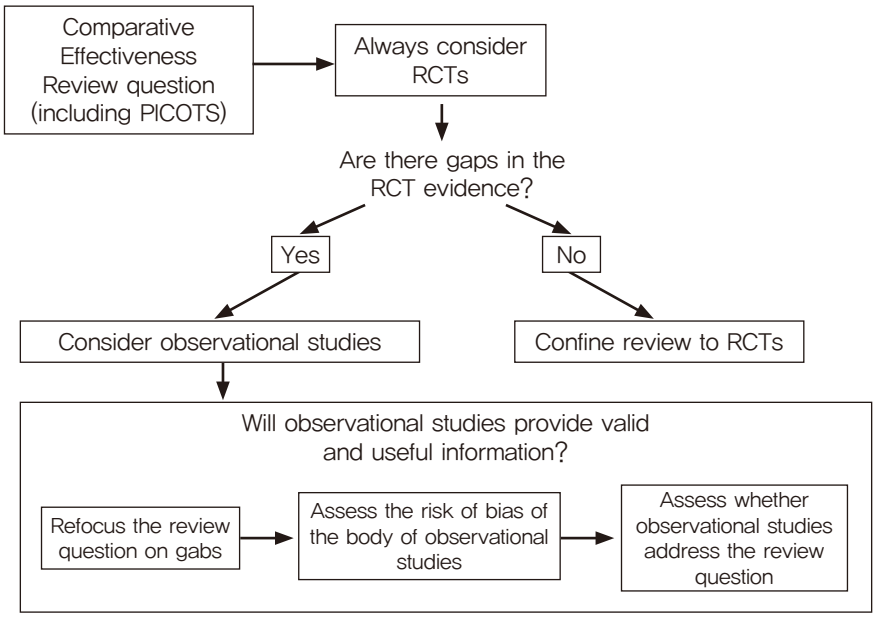

그림 3-3 비교효과 체계적 문헌고찰에서 관찰연구를 포함하는 과정

Reprinted from Observational studies in systemic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program, 64(11), Norris, Susan et al, Journal of Clinical Epidemiology, 1178-86, Copyright (2011), with permission from Elsevier.

이 틀에 의하면 우선 비교효과 체계적 문헌고찰에서 연구하고자 하는 주제에 대해 일차적으로는 RCT를 고려해야한다. 하지만 RCT가 해당 주제에 대해 적절한 답을 제시해주지 못하는 경우가 있다. 예를 들어, 사용하게 될 중재방법이 위험할지도 모르거나 중재법의 효과가 환자의 선호도나 신념에 의존적인 경우에 무작위배정 방법을 사용하는 것은 적절하지 않을 수 있다. 또한 특정 중재방법이 분명히 이득을 보이는 경우에도 RCT는 불필요한 일 일수 있다. 그 외에도 실제 임상연구환경(예를 들어, 약물이 이미 시장에서 판매되고 있을 때 연구대상 모집시 문제가 있는 경우, 중재 결과에 대한 장기간의 추적연구가 필요한 경우, 중재 단위별로 무작위배정이 어려운 경우)에서 RCT는 적용되기가 어려울 수 있다.

해당 연구 질문에 대해 RCT가 적절하지 않다고 판단하였다면, 다음으로 관련 관찰연구가 타당하고 유용한 정보를 제공할지를 평가하여 관찰연구를 포함할지 여부에 대해서 결정해야한다. 이러한 결정을 할 때는 다음의 과정을 따른다. 첫째, RCT로 해결되지 않은 문제를 반영하여 연구 질문을 다시 설정한다. 둘째, 관찰연구의

비뚤림 위험도를 평가한다. 관찰연구에서의 비뚤림 위험도는 아직 합의된 사항은 없으나 다양한 시도들이 있으며, 이에 대해서는 '3.2.3. 비뚤림 위험 평가'에서 기술하기로 한다. 셋째, 최종적으로 포함된 관찰연구들이 연구 질문에 대한 답을 제시해 줄 수 있는지 여부를 판단한다.

위해나 이상반응에 대한 연구에서는 반드시 관찰연구를 포함해야 한다. 아직 위해나 이상반응을 체계적으로 검색하는 가장 좋은 방법이 무엇인지에 대해 실증적인 근거는 없지만, 효과나 효능을 탐색할 때 보다는 더 넓은 포함기준을 가져야하는 것은 확실하다. 관찰연구를 포함하는 경우 문헌분류를 통해 포함된 개별연구를 연구설계별로 분류하여야 비뚤림 위험도를 평가할 수 있다. 문헌분류 작업은 DAMI(-study design algorithm for medical literature of intervention) 도구를 이용한다. 이에 대해서는 「NECA 체계적 문헌고찰 매뉴얼(김수영 등, 2011)」을 참고하기 바란다.

3.2.3. 비뚤림 위험 평가 (문헌의 질평가)

체계적 문헌고찰에서는 비뚤림 위험 혹은 질평가를 수행하는 과정이 필수적이다. "비뚤림 위험(risk of bias)"은 "문헌의 질(study quality)", "방법론적 질(methodological quality)", "연구의 제한점(study limitations)" 또는 "내적 타당도(internal validity)"라는 용어로 표현되기도 하며, 이는 포함된 연구들을 대상으로 해당 연구에서 중요하다고 생각하는 부분을 평가하기 위해 사용하는 도구이다. 일반적으로 "연구의 질"은 비뚤림 위험 보다 더 큰 개념인데, 이것은 중재의 결과측정, 통계처리법, 약물의 용량, 빈도, 강도 등 중재의 형태와 보고 등을 포함한다. "질"이라는 단어는 우연에 기인한 오류(예. 부적합한 표본수 때문에) 또는 잘못된 추론(예. 연구결과의 잘못된 해석)을 모두 포함하는 개념이다(Lohr and Carey, 1999).

비교효과 체계적 문헌고찰에서 비뚤림 위험을 평가하는 방법은 선정된 연구의 디자인에 따라 달라질 수 있다. 여기에서 연구의 디자인 자체가 연구의 비뚤림 위험 평가의 절대적 기준이 되어서는 안 된다. 예를 들어 근거의 계층(evidence hierarchy) 상에서 높은 위치를 차지하고 있는 RCT 형태의 연구라고 해서 무조건 연구의 질이 높은 것은 아니다. 결론적으로 포함된 연구 디자인 별로 비뚤림 위험을 평가해야한다.

관찰연구에서 비뚤림 위험을 평가하는 방법은 아직 합의가 되지 않았다. AHRQ에서는 관찰연구에서 비뚤림 위험을 일으킬만한 요소에 대해 29개의 아이템을 개발하였고, 체계적 문헌고찰 수행시에는 포함된 개별 연구 디자인에 따라 평가해야할

항목을 골라서 평가하도록 하고 있다(Viswanathan and Berkman, 2012). 코크란 비무작위연구 그룹(Cochrane non-randomized study group)에서도 비무작위연구의 비뚤림 위험 평가에 대해 개발하고 있으나 이는 코호트 형태의 연구에만 적용 가능한 형태로 보인다(Reeves 등, 2012) 「NECA 체계적 문헌고찰」에서는 관찰연구 등 비무작위연구의 비뚤림 위험 평가에 대해 RoBANS(Risk of Bias Assessment tool for Non-randomized Studies)를 적용할 것을 권고하였다. 이 도구의 특징은 전후비교를 포함한 비교 관찰연구의 모든 형태에서 적용 가능하다는 특징을 가지고 있다. RoBANS에 대한 자세한 사항은 「NECA 체계적 문헌고찰(김수영 등, 2011)」을 참고하기 바란다.

3.2.4. 근거의 합성 : 정량적 합성, 정성적 합성

체계적 문헌고찰에서 근거를 합성하는데 정량적 합성(quantitative synthesis)을 할지 정성적 합성(qualitative synthesis)을 할지 여부는 미리 계획되어야 하며, 중재의 효과 중 무엇에 대해서 평가할지는 포함된 연구들의 체계적이고 명확한 설명에서부터 시작되어야 한다(CRD, 2009; Deeks 등, 2008). 코크란 핸드북에서는 정량적 합성과 정성적 합성의 특징을 제시하였는데, 첫째, 효과의 방향은 어떠한가? 둘째, 효과의 크기는 어떠한가? 셋째, 효과가 연구들 사이에 일관성이 있는가? 이러한 질문에 대해 통계적 방법을 제시해주는 것이 정량적 합성, 즉 메타분석이다. 이러한 질문에 통계적인 방법을 이용하여 메타분석을 할 수 없으면 정성적인 합성을 해야 한다. 정성적 합성은 통계적인 방법을 통해서 하는 것이 아니라 주관적인 방법을 사용하며 이 경우에도 앞의 질문에 대답할 수 있도록 해야 한다.

임상에 대한 체계적 문헌고찰과 공중 보건 분야에서의 체계적 문헌고찰로 나누어 볼 때 공중 보건 분야에서는 정성적 합성방법을 사용하는 것이 좋다. 대부분의 임상에 대한 체계적 문헌고찰은 RCT를 대상으로 수행하기 때문에 비교적 동질성이 높고 비뚤림 가능성이 비교적 적다. 하지만 공중 보건 분야에서의 효과를 보는 대부분의 연구는 관찰연구인데 이는 바이어스와 교란(confounding)에 의해 영향을 강하게 받기 마련이다. 이 때 메타분석을 하게 되면 내적 타당도에 문제가 생겨 효과를 과대 추정하게 된다. 따라서 공중 보건 분야에서의 체계적 문헌고찰은 정성적 합성을 하고 메타분석은 피하는 것이 좋다.

각 개별연구에서 정성적 혹은 정량적으로 자료 추출을 할 때, 연구들 각각의 방법, 대상, 세팅, 배경, 중재법, 중재결과(outcome), 결과(result), 출판여부, 투자자 등에 관해서 자세히 요약해야 하며, 이하 메타분석에 대한 내용은 '4. 메타분석'을

참고하도록 한다.

3.2.5. 체계적 문헌고찰의 활용

앞서 서술한대로 체계적 문헌고찰은 임상현장과 보건의료분야의 의사결정 등에 있어서 정보를 제공하는 목적으로 수행되며, 체계적 문헌고찰을 통해 연구 질문에 대해 해당 중재의 유효성과 안전성에 대한 과학적인 정보를 제공할 수 있다. 하지만 체계적 문헌고찰을 수행한 결과 선정된 문헌이 부족하거나 연구의 질이 떨어지거나 하는 등의 문제로 근거가 불충분하여 결론을 유보해야하는 경우가 있다. 이런 부분을 연구의 빈틈(research gap)이라고 이야기할 수 있는데, 이는 현재까지 사용가능한 근거로서는 해당 연구 질문에 적절한 결론을 내릴 수 없다는 이야기이다. 이와 같이 체계적 문헌고찰은 현재까지의 근거를 합성할 목적뿐만 아니라 보고자 하는 연구 질문에 대해 적절한 근거가 불충분하다는 연구의 빈틈을 밝히는 목적을 가지고 있다. AHRQ에서는 이러한 연구의 빈틈을 미래 연구 수요(future research needs)로 정리하고 있는데 여기에서는 단순히 향후에 어떤 연구가 필요하다는 수준 외에 연구의 범위, 핵심 질문, 포함해야할 연구 디자인 등을 구체적으로 기술하고 있다. 자세한 내용은 AHRQ의 효과적인 보건의료 프로그램(Effective Health Care Program)내의 미래 연구 수요 프로젝트(Future Research Needs Projects)에서 확인해볼 수 있다[2)]

3.3. 사례

AHRQ에서는 지속적으로 치료, 진단, 스크리닝 등 다양한 비교효과 체계적 문헌고찰의 보고서를 발간하고 있다. 또한 연구 결과를 소비자, 임상의사, 혹은 정책결정자들이 활용할 수 있도록 짧고 평이한 용어로 정리하여 요약본도 함께 제시하고 있다. 뿐만 아니라 새로운 진단이나 치료에 대한 보고서와, 정책연구자들이 참고할 향후 연구 수요에 대한 보고서도 공개하고 있다. 이러한 사례들은 AHRQ 비교효과 체계적 문헌고찰(웹페이지 http://www.ncbi.nlm.nih.gov/books/NBK42934/)에

2) 웹페이지: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=521

서 확인할 수 있다.

한국보건의료연구원에서 수행한 연구 중에서는 「급성 심근경색증 환자에서 약물방출 스텐트와 금속 스텐트의 비교」가 비교효과 체계적 문헌고찰에 해당한다. 이 연구는 급성 심근경색증 환자에서 약물 방출 스텐트(Drug-Eluting Stents, DES)와 금속 스텐트(Bare-Metal Stents, BMS)의 효과 비교에 대한 체계적 문헌고찰을 통해 두 스텐트 간의 임상적 효능 및 안전성을 평가한 연구로, 포함된 연구 디자인으로 RCT와 비무작위시험을 모두 고려하였다. 결과지표로서 사망에 대해서는 RCT 14개의 연구(7,654명)에서 Risk Ratio 0.88(95% CI 0.70-1.11)을 보였고, 비무작위시험 33개(44,849명)에서 0.82(95% CI 0.73-0.91)를 보였다(변량효과모형).

기타 자세한 결과는 NECA 홈페이지의 '연구보고서'에서 「급성 심근경색증 환자에서 약물방출 스텐트와 금속 스텐트의 비교(최동훈 등, 2009)」를 참고하도록 하며, RCT와 비무작위시험의 연구결과를 합성하는 방법으로는 「베이지안 메타분석(장은진 등, 2013)」을 참고하기 바란다.

참고문헌

김수영 등. NECA 체계적 문헌고찰 매뉴얼. 한국보건의료연구원. 2010

AHRQ Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2005–.Available from: http://www.ncbi.nlm.nih.gov/books/NBK42934/

Hartling L, Bond K, Santaguida PL, Viswanathan M, Dryden DM. Testing a tool for the classification

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of InterventionsVersion

Jonathan J Deeks, et al. Cochrane Handbook for Systematic Reviews of Intervention ver. 5.1. Chap 9.

Norris SL, Atkins D, Bruening W, Fox S, Johnson E, Kane R, et al. Observational studies in systemic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011 Nov;64(11):1178–86.

Reeves B, Shea B, Well G. Guidance for includig non–randomized studies (NRS) in Cochrane reviews: conclusions from an invited workshop. 67. Abstracts of the 20th Cochrane Colloquium, 30 September–3 October 2012, Auckland, New Zealand. Cochrane Database of Systematic Reviews, Supplement 2012: Art. No. CD201200.DOI:10.1002/14651858.CD201200.

Robinson KA, Saldanha IJ, Mckoy NA. Frameworks for Determining Research Gaps During Systematic Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Jun. (Methods Future Research Needs Reports, No. 2.) Available from: http://www.ncbi.nlm.nih.gov/books/NBK62478/

Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health. 2012 Apr;33:425–45.

Suh HS, Song HJ, Choi JE, Jang EJ, Son HJ et al. Drug–eluting stents versus bare–metal stents in acute myocardial infarction: A systematic review and meta–analysis. Int J Technol Assess Health Care. 2011;27(1):11–22.

Talati R, Scholle JM, Phung OJ, et al. Effectiveness and Safety of Antiepileptic Medications in Patients With Epilepsy [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Dec. (Comparative Effectiveness Reviews, No. 40.) Introduction. Available from: http://www.ncbi.nlm.nih.gov/books/NBK83926/

Tim Carey et al. Prioritizing Future Re샘search through Examination of Research Gaps in Systematic Reviews. Patient-Centered Outcomes Research Institute (www.pcori.org). 2012

Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol. 2012;65(2):163-78.

PART 4

메타분석

비교효과연구에서는 관심 연구주제에 대한 기존 연구가 충분히 있을 경우 체계적 문헌고찰을 실시하고, 체계적 문헌고찰 결과 최종 선정된 연구들을 대상으로 양적 근거합성을 하고자 할 경우 메타분석(meta-analysis)을 실시하게 된다. 만일 관심 있는 두 군의 직접비교 연구가 충분히 있다면 두 군을 직접 비교하는 전통적 메타분석 또는 베이지안 메타분석을 실시할 수 있다. 하지만 관심있는 두 군의 직접비교 연구가 충분하지 않은 경우 간접비교(Indirect Treatment Comparison, ITC)를 실시할 수 있으며, 비교군이 여러 개인 경우는 혼합비교(Mixed Treatment Comparison, Multiple Treatment Comparison, MTC)를 실시할 수 있다. 보통 간접비교와 혼합비교 방법을 통틀어서 네트워크 메타 분석(network meta-analysis)이라고 한다. 간접비교와 혼합비교도 전통적 방법과 베이지안 방법 모두 적용이 가능하다. 이 장에서는 직접비교 방법과 간접비교 및 혼합비교 방법을 포함하는 네트워크 메타 분석에 대해 다루고자 한다.

4.1. 전통적 메타분석[3)]

4.1.1. 개요

메타분석은 2개 이상의 개별 연구 결과들을 합성하여 통합된 가중평균 요약 추정치를 정량적으로 산출함으로써 해당 중재법의 임상적 효과를 평가하는 통계적 기법이다. 일반적으로 체계적 문헌고찰에서 결과를 정량적으로 분석할 때 메타분석을 주로 이용하지만, 반드시 메타분석을 포함해야만 체계적 문헌고찰이 되는 것은 아니다. 메타분석은 연구 대상군, 중재법, 중재결과가 충분히 동질한 경우에만 수행되어야 한다. 메타분석을 수행하는 이유는 첫째, 검정력을 증가시키기 위함이다. 여러 연구가 통합되면 표본크기 역시 증가하기 때문에 효과크기가 작은 경우도 발견할 기회가 높아지게 된다. 둘째, 정밀성을 향상시키기 위함이다. 중재법에 대한 효과추정치는 보다 많은 정보가 있을 경우 정밀도가 향상될 수 있다. 셋째, 개별 연구들이 조사하지 않은 질문에 답하기 위해 메타분석을 수행한다. 즉, 개별 연구 간 중재 효과의 일관성을 조사할 수 있으며, 조사된 효과추정치에 차이가 있는 경우 이에 대한 설명이 가능하다. 넷째, 상반된 연구들 사이에서 발생하는 논쟁을 조정하거나 새로운 가

3) 김수영 등(2011), NECA 체계적 문헌고찰 매뉴얼.

설을 만들기 위함이다. 결과에 대한 통계적 분석을 통해 상반된 정도를 평가할 수 있으며, 개별 연구들 사이에서 나타나는 결과 차이에 대한 원인을 분석하여 새로운 가설을 제시할 수 있다. 반면, 다음과 같은 경우는 메타분석을 수행하면 안 된다. 첫째, 대상군이나 중재법과 같은 연구의 특성이 다른 개별 연구들이 포함된 경우이다. 둘째, 포함되는 개별 연구들에 바이어스의 위험이 있는 경우이다. 마지막으로, 심각한 보고 바이어스(reporting bias)가 있는 경우이다. 메타분석 시 특정한 특성을 가지거나 특정한 방향의 연구 결과만이 포함되면 부적절한 요약추정치를 얻게 된다.

4.1.2. 방법

4.1.2.1. 분석 단계

메타 분석의 분석 단계는 크게 다음과 같이 나누어 볼 수 있다.

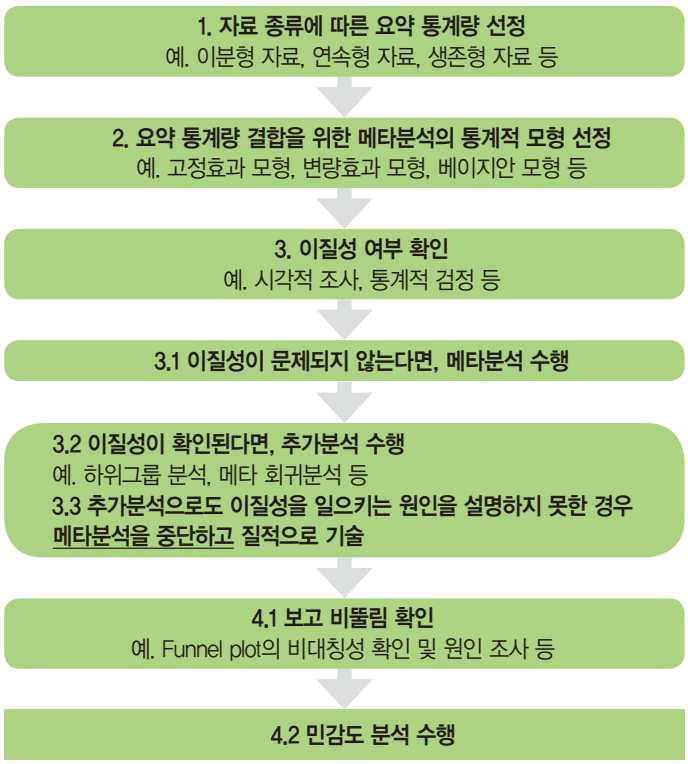

그림 4-1 메타분석의 절차

(자료원 : 김수영 등(2011), NECA 체계적 문헌고찰 매뉴얼)

(1) 자료 종류에 따른 요약 통계량 선정

메타분석에서 사용할 수 있는 자료의 종류로 이분형 자료(dichotomous data), 연속형 자료(continuous data), 생존형 자료(time-to-event data)를 들 수 있다. 이분형 자료는 임상연구에서 가장 일반적으로 사용되며 (예. 사망 또는 생존), 이분형 자료 연구에서 효과크기는 다음 네 가지 중 한가지로 보고될 수 있다.

- 상대위험도(risk ratio 또는 relative risk): 두 군 간의 사건이 발생할 확률에 대한 비율
- 오즈비(odds ratio): 두 군 간의 사건의 오즈에 대한 비율
- 위험차(risk difference): 중재군에서의 사건이 발생할 위험과 대조군에서의 사건이 발생할 위험의 차이
- 필요치료수(number needed to treat, NNT): 해당기간에서 특정 사건의 위해를 피하거나 편익을 발생시키기 위해 비교 중재법에 비해 시험 중재법을 더 받아야하는 사람의 기대 수

연속형 자료는 중재나 결과가 연속적인 수치로 보고되거나 순서형 자료이지만 범주의 개수가 많은 경우를 일컫는다. 일반적으로 자료가 대칭적인 분포를 이루고 있다고 가정하기 때문에 추출된 자료의 분포가 치우쳐 있는 경우 로그변환이 필요하다. 연속형 자료에서 주로 사용하는 요약 통계량으로는 평균차와 표준화된 평균차가 있다.

- 평균차(mean difference): 동일한 측정 도구로 측정되었을 때 두 군 간의 평균값의 차이
- 표준화된 평균차(standardized mean difference): 다양한 측정 도구로 결과를 측정하였을 경우, 단일 단위로 연구결과를 표준화하기 위한 요약 통계량으로 평균차를 연구대상자들에서 관찰된 표준편차로 나눈 값

생존형 자료는 사건이 발생하기 전의 시간경과에 초점을 둔 자료로 각 연구 참여자는 관찰기간 동안 관심 사건이 관찰되지 않았거나(censored 자료) 관찰하고자 했던 사건이 완전하게 관찰될 수 있었던 경우(uncensored 자료)로 분류된다. 생존형 자료에서 사용되는 요약 통계량은 위험도비이다.

- 위험도비(hazard ratio): 대조군에서 사건이 발생할 위험과 중재군에서 사건이

발생할 위험 또는 확률에 대한 비

(2) 요약 통계량 결합을 위한 메타분석의 통계적 모형 산정

메타분석의 수행은 우선, 개별 연구의 요약 추정치(summary estimate)를 계산하는데 이때 정보 제공의 정도에 따라 각각의 연구에 가중치를 부여하게 된다. 일반적으로 표본크기가 가중치로 사용되지만, 사건 발생률이나 분산의 역수, 연구의 질 등이 사용되기도 한다. 다음, 가중치가 부여된 평균(weighted average)을 계산하여 통합 요약 통계량를 산출한다. 요약 통계량을 결합하기 위한 메타분석의 통계적 모형에는 고정효과 모형(fixed-effects model)과 변량효과 모형(random-effects model)이 있다.

고정효과 모형은 각 연구들에 있어 중재효과의 참값은 하나라는 전제 하에 관찰된 치료효과 값들의 차이는 표본추출의 오차(sampling error) 때문이라는 가정에서 출발하는 모형이다. 고정효과 모형을 고려할 수 있는 경우로는 우선, 메타분석에 통합된 연구들의 인구집단이나 중재법 등이 동질하다고 판단되는 경우, 둘째, 메타분석의 목적이 하나의 진실된 효과크기를 추정하기 위한 경우, 셋째, 메타분석에 통합된 연구들의 수가 매우 적은 경우이다. 연구들의 수가 적으면 연구간 변동(between-study variance)을 추정할 수 없기 때문에 고정효과 모형만이 사용가능하다. 고정효과 모형에서 가중치를 추정하는 방법은 역분산(generic inverse variance)추정법, 멘텔-헨젤(Mantel-Haenszel) 추정법 등이 있다.

- 역-분산 추정법: 효과 추정치 분산의 역수를 개별 연구의 가중치로 사용하는 방법. 이를 이용하면 작은 표준 오차를 가지는 큰 규모의 연구는 더 많은 가중치를 받게 된다. 역-분산 방법은 결합하고자 하는 연구들의 수는 적지만 각 연구들의 표본수가 큰 경우에 효과적이다.
- 멘텔-헨젤 추정법: 오즈비나 상대위험도와 같은 2×2 표를 만들 수 있는 자료에서 사용되며, 사건 발생률이 낮거나 연구의 규모가 작지만 메타분석에 포함되는 연구의 수가 많을 경우 효과적인 방법이다. 사건 발생률이 높은 경우는 역-분산 방법과 유사한 통합된 치료효과 크기를 추정하며, 처리 효과들을 결합할 때 각 연구의 오즈비나 상대위험도에 대해 로그변환을 하지 않고 오즈비값 자체를 사용한다.

변량효과 모형은 중재효과의 참값은 하나만 존재하는 것이 아니라 각 연구들에 있어 중재법의 효과는 중재효과의 평균 참값을 중심으로 정규분포를 따른다고 가정

하는 모형이다. 따라서 각 연구의 중재효과들 간에 관찰되는 변동(variation)은 표본추출의 오차와 각 연구들 간의 변동이 함께 표현된 것이라고 가정한다. 연구 간의 이질성이 심한 경우 변량효과 모형은 고정효과 모형에 비해 치료효과의 유의성에 대해 더 보수적인 추정치(더 넓은 신뢰구간)를 제공한다. 변량효과 모형은 고정효과 모형에 비해 작은 규모의 연구에 더 높은 가중치를 주고 더 큰 규모의 연구에는 더 낮은 가중치를 주기 때문에 치료효과의 추정치는 두 모형 간에 다를 수 있다. 그러나 작은 규모 연구들의 결과가 큰 규모 연구들의 결과와 체계적으로 다를 경우 정규성의 가정이 충족되지 못하므로 변량효과 모형은 적절한 효과 추정치를 제공하지 못한다. 변량효과 모형을 고려할 수 있는 경우는 첫째, 메타분석에 포함된 연구의 인구집단이나 중재법 등이 동일하지 않거나 둘째, 메타분석의 목적이 폭넓은 인구집단에서의 일반화일 때이다. 변량효과 모형에서의 가중치는 고정효과 모형과는 달리 연구간 변이(between-study variance)가 추가되며, 가중치 추정시 일반적으로 사용되는 방법으로 DerSimonian and Laird 추정법이 있다. 변량효과 모형이 연구간 변동을 고려하지만 이질성 문제 해결을 위한 이상적인 방법은 아니므로 연구간 이질성이 심할 경우 변량효과 모형으로부터 계산된 효과 추정치는 부적절할 수 있다.

(3) 이질성 여부 확인

메타분석에서의 이질성(heterogeneity)은 메타분석에 포함되는 개별 연구들의 결과 간 변동이 표본추출 오차 이상으로 관찰되어 우연으로 설명할 수 없는 것을 뜻한다. 이질성을 야기시키는 요인으로는 임상적 또는 방법론적 다양성, 우연, 그리고 바이어스를 들 수 있다. 임상적 다양성은 연구된 환자나 인구집단의 다양성이나 사용된 중재법, 연구들 간의 결과를 측정하기 위해 사용된 요약 추정치, 또는 추적조사 기간의 다양성을 의미한다. 이에 반해, 방법론적 다양성은 연구유형과 바이어스 위험의 다양성이나 치료효과 크기의 방향과 크기의 다양성을 말한다. 이질성의 확인 및 평가는 그래프를 이용한 시각적 조사나 통계적인 검정 방법을 통해 이루어진다.

그래프를 이용하여 이질성을 평가하는 방법으로는 숲그림(forest plot)과 L'Abbe plot 등이 있다. 숲그림은 개별 연구들의 요약 추정치와 신뢰구간을 나열한 그림으로, 이를 통해 개별 연구들의 치료 효과값의 방향성과 신뢰구간이 겹치는지 여부를 확인할 수 있다. 신뢰구간이 겹치지 않거나 적은 정도로만 겹친다면 추가적으로 통계적 검정 방법을 수행한다.

L'Abbe plot은 대조군에서의 사건 발생률에 대한 중재군의 사건 발생률을 제시

하는 그림이며 이분형 자료에서 특히 유용하다. 그래프에서 원의 크기는 표본의 크기이며 연구간 이질성이 심할 경우 원들은 회귀선 주변에 매우 퍼져있게 된다. 주로 대조군과 중재군 또는 두 군 모두 중 어떤 군이 변동의 원인인지 판단하는데 도움을 주지만 이질성 여부 판정에 흔히 쓰이지는 않는다.

이질성을 평가하는 통계적 검정 방법으로 카이제곱 검정법(Q statistics)과 Higgin's I^2statistic을 들 수 있다. 카이제곱 검정은 각 연구들의 중재효과가 공통중재효과값으로부터 얼마나 멀리 떨어져 있는지를 검정하는 방법이다. 카이제곱 검정은 메타분석에 포함된 연구의 규모가 작거나 포함된 연구의 수가 적을 경우 낮은 검정력을 가지며, 반대로 포함된 연구의 수가 많을 때는 불필요한 이질성을 발견할 수 있기 때문에 해석에 주의를 기울여야 한다. 따라서 통계적 유의수준을 0.05 대신 0.10으로 사용한다. 카이제곱 검정법에서 귀무가설은 이질성이 있다는 근거를 찾을 수 없다이며 귀무가설 채택시 고정효과 메타분석을 수행하며 귀무가설 기각시 변량효과 메타분석이 고려된다. 그러나 귀무가설이 채택된 경우라도 포함된 연구들의 특성이 동질하지 않거나 메타분석의 목적이 일반적으로 폭넓은 인구집단에서 일반화시키는 것일 때 변량효과 모형이 고려될 수 있다.

Higgin's I^2통계량은 비일관성을 정량화시킨 통계량이며 I^2값이 40%이하이면 이질성이 중요하지 않을 수 있음으로, I^2값이 75% 이상이면 상당한 이질성이 있음으로 해석된다.

이질성에 대한 통계적 검정력은 포함된 연구의 수와 개별 연구에 배정된 가중치에 따라 달라지며 특히 한 연구가 전체 정보의 대부분을 제공할 때 낮아진다. 따라서 포함된 연구가 10개 미만인 경우 이질성에 대한 통계적 검정 방법으로 연구들 간에 존재하는 결과의 이질성을 발견하지 못할 수 있다.

(3-1) 이질성이 문제되지 않는다면, 메타분석 수행

이질성이 문제되지 않는다면 통합된 요약추정치를 구하기 위한 메타분석 방법을 각 모형의 기본 가정을 고려하여 결정한다. 2011년 한국보건의료연구원에서 발간한 체계적 문헌고찰 매뉴얼에서 메타분석의 방법에 대해 자세히 기술하고 있으므로 이를 참고한다.

(3-2) 이질성이 확인된다면 추가분석 수행

연구들 간의 이질성에 대한 원인을 파악하기 위해 하위그룹 분석이나 메타 회귀분석(meta-regression)과 같은 추가 분석을 수행할 수 있다. 하위그룹 분석은 연구수준의 변수 (예. 연구설계, 연구의 질, 질환의 중증도 등)를 사전에 구체화한 후

에 수행한다. 구체화된 특성들에서 하위그룹의 수를 제한하여 통계적인 오류의 가능성을 줄여야하며, 하위그룹이 분석되는데 영향을 미치는 연구 결과들을 연구자가 모르게 하여 바이어스가 있는 결과를 도출할 가능성을 줄여야 한다.

메타 회귀분석은 연구수준의 공변량에 대한 연구 요약 추정치의 회귀분석이다. 따라서 분석단위는 개별 연구이며 치료효과와 조사된 연구 특성 간에 선형관계가 있다면 메타 회귀분석은 하위그룹 분석보다 통계적 검정력이 더 크다. 회귀분석에서 종속변수는 로그변환된 오즈비나 표준화된 평균차 등과 같은 중재효과 크기이며, 설명변수는 중재효과 크기에 영향을 미칠 수 있는 연구들의 특성 (예. 연구대상자의 특성, 임상상황, 추적관찰 기간 등)이다. 연구들 간의 모든 변동은 회귀모형에 포함되는 공변량으로 설명 가능하다고 판단될 때 고정효과 메타 회귀분석이 적절하다. 반면에, 변량효과 메타 회귀분석은 공변량으로 설명되지 않는 연구들 간의 변동인 잔여 이질성(residual heterogeneity)을 허용한다.

(4-1) 보고 바이어스 확인

보고 바이어스는 연구 결과의 확산이 결과의 특성과 방향의 영향을 받을 때 발생하는 바이어스로 출판 바이어스(publication bias)가 대표적인 예이다. 출판 바이어스는 연구 결과의 통계적 유의성과 출판 가능성 사이에 관련성이 있을 경우 나타나는 바이어스이며 통계적으로 유의한 긍정적인 연구 결과의 경우 출판이 상대적으로 용이할 수 있기 때문에 발생한다. 보고 바이어스 중 출판 바이어스에 특히 주의를 기울여야 하는 이유는 메타분석 시 출판된 연구들의 결과만을 통합할 경우 치료효과를 과대 추정할 위험이 있기 때문이다. 보고 바이어스를 확인하는 방법에는 그래프를 이용한 방법과 통계적 검정 방법이 있다.

보고 바이어스를 확인하기 위해 사용되는 그래프를 이용한 방법으로 깔대기 그림(funnel plot)과 Contour-enhanced funnel plot을 들 수 있다. 깔대기그림은 연구의 표본크기(세로축)에 대한 개별 연구들의 추정된 효과크기(가로축)를 제시하는 산포도이다. 시각적으로 깔대기그림이 대칭이면 보고 바이어스의 가능성이 낮다고 간주한다. 대체로 작은 규모의 연구들은 그래프의 하단에 넓게 분포되고, 큰 규모의 연구들은 깔대기 윗부분의 좁은 부분에 분포하게 된다. 깔대기그림 사용 시 유의할 점은 그래프의 비대칭성 여부는 적어도 10개 이상의 연구가 포함된 경우에만 해석이 가능하며, 깔대기그림의 비대칭성이 반드시 보고 바이어가 있음을 의미하는 것은 아니라는 것이다. 깔대기그림의 비대칭성을 야기하는 요인으로는 출판 바이어스를 포함한 선택 바이어스, 낮은 수준의 연구의 질, 이질성 및 우연을 들 수 있다.

Contour-enhanced funnel plot은 깔대기그림에 통계적 유의성을 추가로 제시하여 출판 바이어스와 비대칭성의 다른 원인들을 구분할 수 있도록 돕는다. 연구들이 통계적으로 유의하지 않은 영역에서 결측되었다면 깔대기그림의 비대칭성은 출판 바이어스 때문일 가능성이 있지만, 통계적으로 유의한 영역에서 결측되었다면 출판 바이어스가 비대칭의 원인일 가능성은 낮아진다. 따라서 메타분석에서 Contour-enhanced funnel plot을 수행할 것을 권장한다.

깔대기그림의 비대칭 정도가 통계적으로 유의한지 여부를 판단하기 위해 Begg와 Mazumdar의 검정과 Egger의 검정을 사용할 수 있다. Begg와 Mazumdar의 검정(이하 Begg의 검정)은 표준화된 중재 효과값과 해당 표준오차 간의 상관관계를 평가하기 위한 보정된 순위상관(rank correlation) 검정으로 개별 연구의 효과크기와 표본크기가 작은 경우 검정력이 증가된다. 그러나 메타분석에 포함된 연구의 수가 적다면 결과 해석 시 주의를 기울여야하며, 특히 검정결과가 통계적으로 유의하지 않더라도 바이어스의 가능성은 배제될 수 없다. Egger의 검정은 중재 효과값의 역수에 의해 가중치가 부여된 표준오차에 대한 중재효과 추정치의 선형 회귀검정으로 Y절편이 0과 유의하게 다른지 평가한다. Egger의 검정은 소규모 연구 효과 (즉, 작은 규모의 연구들이 더 큰 치료효과 크기를 보이는 경향성)를 검정하는데 사용되며, 중재효과와 표준오차 사이에 연관성이 클수록 기울기는 0에서 멀어지게 된다.

(4-2) 민감도 분석 수행

민감도 분석은 메타분석 연구결과의 강건함을 확인하기 위해 수행되는 분석 방법이다. 가정이 달라짐에 따라 연구결과가 달라진다면 그 원인을 찾고 다른 가정에 따라 민감도 분석을 수행하는 등의 적절한 조치를 취해야한다. 그 이유가 연구 특성에 따른 것이라면 연구 특성들에 따라 일차 연구들을 포함 혹은 배제시켜 분석하고 결과의 차이를 비교해볼 수 있다. 서로 분석결과가 다른 경우 높은 수준의 질을 가진 연구의 통합 추정치에 초점을 두어 결론을 내려야 한다. 또한, 결과에 대한 이질성의 원인이 예외적인 연구들 때문이라 판단되면, 민감도 분석의 일부로서 이상치를 포함 혹은 배제시켜 결과를 관찰해볼 수 있다.

4.1.3. 사례

4.1.3.1. 스텐트

한국보건의료연구원은 2009년 급성심근경색증 환자 중 ST분절상승(STEMI)을

가진 환자를 대상으로 약물 방출 스텐트(drug-eluting stents, DES)와 금속 스텐트(bare-Metal stents, BMS)의 효과를 비교하기 위한 메타분석을 시행하였다. 2009년 9월까지 발표된 임상시험(RCT)이나 관찰연구 논문을 바탕으로 DES와 BMS 방법이 STEMI 환자들의 사망, 심근경색 재발, 표적혈관 재관류술(target vessel revascularization, TVR) 그리고 스텐트 혈전증(stent thrombosis, ST)에 미치는 영향을 평가하였다. 연구 방법에 따른 차이로 인해 임상연구(14개)와 관찰연구(35개)를 별도로 분석하였으며, 개별 연구의 결과들은 역-분산 변량효과 모형(inverse-variance random-effects model)을 사용하여 정량적으로 산출되었다. 숲그림(forest plot)과 I^2값을 통해 연구 간 이질성이 평가되었으며, 보고 바이어스(reporting bias)의 유무를 확인하기 위해 깔대기그림과 Egger 검정법 등이 사용되었다. 임상연구들에 대한 메타분석 결과, BMS와 비교하여 DES는 STEMI 환자들의 TVR을 약 50%(RR=0.48, 95%CI=0.41-0.56), 심근경색 재발률은 약 25%(RR=0.77, 95%CI=0.61-0.97) 낮추는 효과를 보였다. 관찰연구들에 대한 메타분석 결과에 따르면, 추적기간을 고려하지 않았을 때 DES는 STEMI 환자들의 사망률을 15%(RR=0.85, 95%CI=0.79-0.91), TVR의 발생률은 약 40%(RR=0.61, 95%CI=0.48-0.77) 감소시켰다(최동훈 등, 2009).

4.1.3.2. 비파열 뇌동맥류

한국보건의료연구원은 2010년 비파열 뇌동맥류의 예방적 치료 성과에 대한 메타분석 연구를 진행하였다. 2010년까지 발표된 논문을 바탕으로 비파열 뇌동맥류 환자를 대상으로 시행된 결찰술(clipping)과 코일 색전술(coiling)의 임상적 결과를 비교하였으며, 개별 연구들의 질과 제시된 증거에 대한 평가는 MINORS와 GRADE를 통해 시행되었다. 개별 연구들의 오즈비가 역-분산법을 이용하여 정량적으로 산출되었으며, 코크란의 카이제곱 검정법과 I2값을 바탕으로 이질성에 대한 평가가 이루어졌다. 총 24개의 개별 연구가 메타분석에 포함되었으며, 분석 결과 색전술에 비해 결찰술에서 장애율이 더 높게 나타났다(OR=2.38, 95%CI=1.33-4.26). 그러나 총사망률과 병원 내 사망률은 두 치료군 간 차이가 나타나지 않았다(이현주 등, 2010).

4.2. 베이지안 메타분석

베이지안 메타분석법은 베이지안 통계를 이용해 메타분석을 수행하는 방법을 말한다. 베이지안 통계에서는 자료의 정보와 사전(prior) 정보를 결합하여 모수에 대한 정보를 업데이트하여 사후분포(posterior distribution)로 표현하며, 이 사후분포를 이용하여 메타분석의 결과값을 추론하게 된다.

베이지안 메타분석에서는 모든 모수의 불확실성을 모형에서 고려할 수 있으며, 외부의 타당한 정보를 포함할 수 있고 모형을 복잡하게 확장하는 것도 가능하다. 또한 메타분석에서 개별 연구의 치료효과를 추정할 때 다른 연구들로부터 정보를 빌려올 수 있으며, 관심있는 모수에 대해 확률로서 표현할 수 있다는 장점이 있다. 베이지안 메타분석의 장점과 단점을 정리하면 〈표 4-1〉과 같다. 구체적인 베이지안 메타분석 방법은 한국보건의료연구원에서 발행한 「베이지안 메타분석(장은진 등, 2013)」을 참고하도록 한다.

표 4-1 베이지안 메타분석의 장 · 단점

장점

① 전문가 의견이나 메타분석 대상에 포함되지 않은 연구의 결과를 사전분포로 반영할 수 있음
② 결과발생이 드물거나 이상반응에 대한 메타분석 시 연속성 수정(continuity correction)을 사용하지 않고 메타분석이 가능함
③ 연구설계에 대한 신뢰성을 사전분포로 반영하거나, 계층적 모형을 사용하여 연구설계가 다른 연구들을 통합할 수 있음
④ 신용구간(credible interval) 계산시 정규분포 가정이 필요 없음
⑤ 통합대상 연구의 수가 작은 경우 유용함
⑥ 관심있는 사건에 대해 확률적으로 언급할 수 있음
⑦ 기존 정보가 주어져 있을 때, 불확실성을 반영하여 예측이 가능함
⑧ 통합 추정치 산출시, 각 연구들의 추정치는 다른 연구들로부터 정보를 빌려와서(borrow strength) 추정됨
⑨ 모든 모수에 불확실성을 허용할 수 있음

단점

① 사후분포 추정이 어려움
② 사전분포에 따라 결과가 민감할 수 있음
③ 계산이 복잡함
④ 소프트웨어가 제한적임

4.3. 간접비교 및 혼합비교

4.3.1. 개요

비교효과연구에서 관심있는 두 군간 효과를 비교하고자 할 경우, 관심있는 두 군간 직접비교한 무작위배정 비교임상시험들이 있다면 가장 이상적이다. 하지만 직접비교 RCT가 없는 경우도 많이 있다. 예를 들어 허가 승인을 위한 임상시험에서 신약은 대부분 위약(placebo) 또는 표준치료제(standard care)와 비교하게 되며, 활성대조군(active control treatment)과는 거의 비교를 하지 않는 경향이 있다. 또한 국가마다 관심있는 대조군이 다를 수 있으므로 동일한 적응증에 대해 많은 치료법을 하나의 임상시험에서 모두 고려하는 것은 실제로 불가능하다. 따라서 관심있는 두 군간 직접비교가 없는 경우, 간접비교와 혼합비교 방법은 최선의 치료를 선택하기 위한 유용한 방법이다.

비교효과연구에서 관심있는 치료 A와 B를 비교하고자 할 경우, 치료 A와 B의 직접비교가 없고 치료 A와 치료 C의 직접비교와 치료 B와 치료 C의 직접비교만 있는 경우, 치료 C를 이용하여 치료 A와 치료 B의 상대치료효과를 추정하는 방법을 공통대조군 간접비교라고 한다. 여기서 치료 C를 공통대조군(common comparator)이라고 하며, 공통대조군을 이용하여 기저상태가 다름에 대한 '일종'의 보정을 하므로 보정된(adjusted) 간접비교라고 한다(그림 4-2).

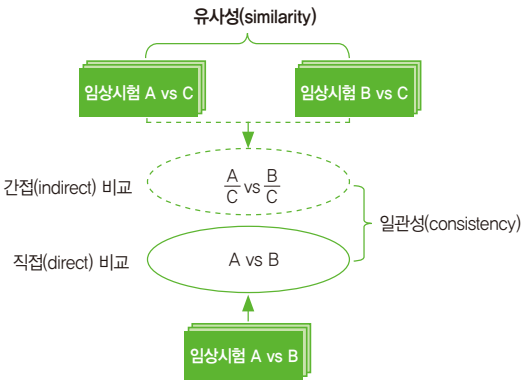

그림 4-2 직접비교와 간접비교

Reproduced from What is Indirect Comparison?, Song, What is...? series, Copyright (2009), with the permission from Hayward Medical Communications.

직접비교가 불가능할 경우에만 간접비교 방법을 이용하는 것이 아니라, 직접비교와 간접비교가 모두 있는 경우 이들을 함께 이용할 수도 있다. 이렇게 직접비교와 간접비교를 동시에 이용하여 효과를 비교하는 방법을 혼합비교라고 한다. 직접비교를 통한 근거가 분명하더라도 혼합비교를 통해 간접비교 자료를 이용하면 직접비교만 이용했을 경우보다 효과 추정치의 정도(precision)가 높아지고, 네트워크 내 포함되어 있는 모든 대상자들을 포함하게 됨으로써 일반화 가능성이 커진다(Lu and Ades, 2004).

네트워크 메타분석에서는 허가를 위한 무작위배정 비교임상시험 뿐만 아니라 실제 임상현장에서 진행되는 실용임상시험까지도 포함하여 이들을 구분하지 않고 효과를 추정하므로 비교효과연구 분야에서 유용한 방법론이라고 할 수 있다. 네트워크 메타분석은 크게 공통대조군 간접비교와 혼합비교로 구분할 수 있으며, 두 방법 간 가장 큰 차이점은 혼합비교의 경우 직접비교가 포함되어 네트워크가 닫힌 형태(closed loop)라는 것이다. 간접비교의 경우 치료군 쌍에 대해 단계적으로 치료효과를 추정할 수 있으며, 간접비교와 혼합비교 모두 공변량을 추가적으로 고려할 수 있고, 전통적 방법인 빈도론자(frequentist) 방법 및 베이지안 방법 모두 적용 가능하다(그림 4-3).

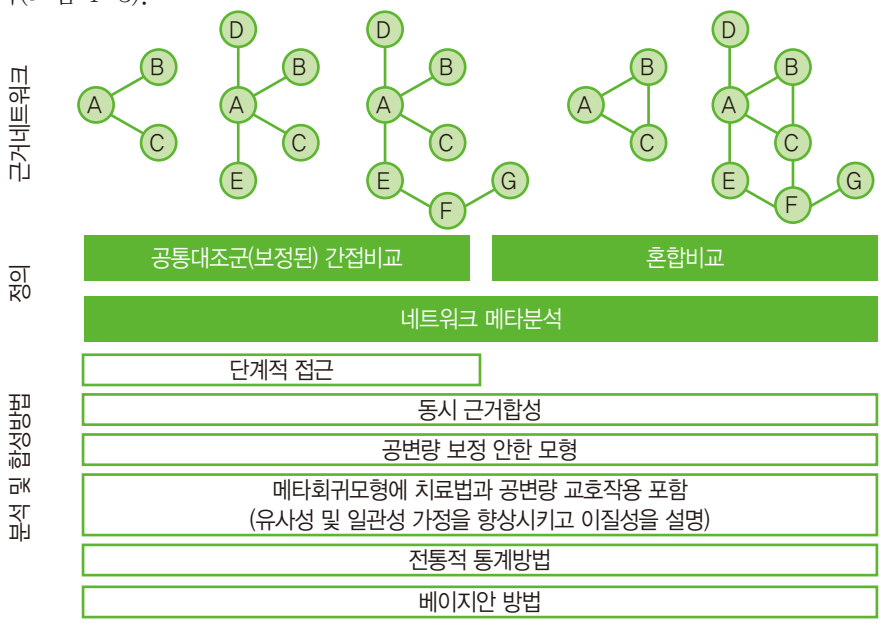

그림 4-3 네트워크 메타분석

Reprinted from Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1, 14(4), Jansen et al, Value in Health, 417-28, Copyright (2011), with permission from Elsevier.

네트워크 메타분석을 수행할 경우 가장 기본적인 조건은 직접비교의 메타분석과 마찬가지로 동질성(homogeneity)이 성립해야 한다. 따라서 치료 A와 C를 비교하는 무작위배정 비교임상시험들이 메타분석으로 통합할 수 있을 만큼 동질적이고, 치료 B와 C를 비교하는 무작위배정 비교임상시험들이 메타분석으로 통합할 수 있을 만큼 동질적이어야 한다. 공통대조군을 이용한 간접비교인 경우 기본적으로 만족해야 하는 동질성 가정 이외 추가적으로 유사성(similarity) 가정을 만족해야 한다. 유사성 가정은 치료 A와 C를 비교하는 무작위배정 비교임상시험과 치료 B와 C를 비교하는 무작위배정 비교임상시험에서 치료 A와 치료 B를 바꾸어서 임상시험을 수행하더라도 기존의 임상시험에서 얻은 것과 동일한 효과를 얻을 수 있다는 가정으로 공통대조군 간접비교에서 반드시 만족해야 하는 중요한 가정이다. 즉 유사성 가정은 치료들을 바꿔 시행하더라도 결과가 동일할 만큼 간접비교로 고려하는 임상시험들의 모집단, 임상시험 조건 등이 유사하다는 의미이다. 유사성 가정은 때로 교환가능성(exchangeability) 가정이라고 표현하기도 한다. 혼합비교는 직접비교와 간접비교를 동시에 이용하는 방법으로, 기본적으로 위에서 언급한 동질성 가정 및 유사성 가정을 만족해야 한다. 여기에 추가적으로 직접비교에 의한 치료효과와 간접비교에 의한 치료효과가 동일하다는 일관성 조건을 만족해야 한다.

위에서 설명한 직접비교에서의 동질성 가정, 공통대조군 간접비교에서 동질성 가정 및 유사성 가정, 혼합비교에서 동질성 가정, 유사성 가정 및 일관성 가정에 대해 요약하면 아래 그림과 같다.

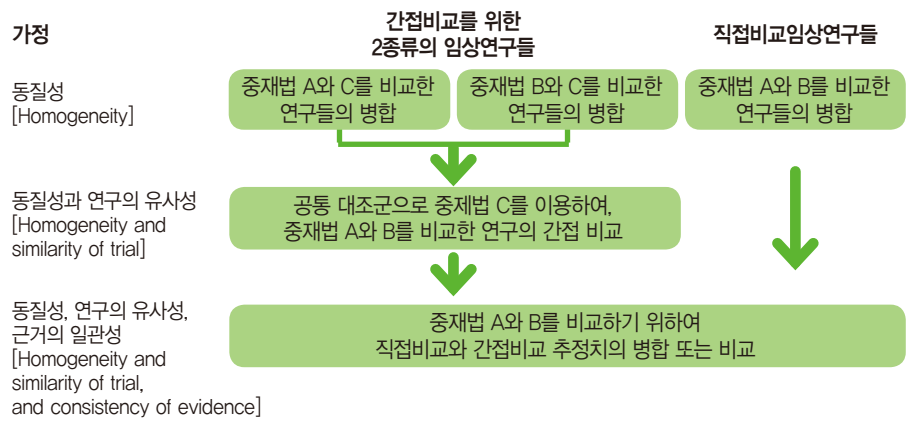

그림 4-4 공통대조군 간접비교와 혼합비교의 가정

4.3.2. 방법

4.3.2.1. 근거 네트워크 작성

간접비교와 혼합비교를 포함한 네트워크 메타분석을 수행하기 위해서는 먼저 그림 4-3와 같이 비교할 치료들에 대한 근거 네트워크(network of evidence) 구조를 작성해야 한다. 근거 네트워크에서 노드(node)는 치료를 나타내며 연결선은 하나 이상의 직접비교 RCT가 있음을 나타낸다.

네트워크 메타분석에서 문헌검색은 비교하고자 하는 치료군이 하나라도 포함된 모든 문헌을 검색해야 하는데, 예를 들어 치료 B와 C를 직접 비교한 연구를 검색한 결과 직접비교 연구가 부족할 경우 네트워크 메타분석을 고려하게 된다. 이 경우 치료 B와 다른 치료에 대한 연구와 치료 C와 다른 치료에 대한 연구를 모두 검색해야 한다. 검색 결과 공통대조군 A가 있는 경우 이를 이용하여 공통대조군 간접비교를 고려해 볼 수 있다. 하지만 적절한 공통대조군이 없을 경우 관심있는 치료들과 비교 가능한 치료들을 규명해야 하는데, 이를 위한 검색은 시간과 노력을 많이 필요로 한다.

4.3.2.2. 가정에 대한 탐색적 검토

치료효과에 영향을 줄 수 있는 연구들의 기저특성, 연구방법론 등에 대한 검토를 통해 기본적으로 메타분석을 수행하기 위해 만족해야 하는 동질성 가정에 대해 먼저 검토하고, 동질성을 만족할 경우 공통대조군 간접비교를 수행할 경우 유사성 가정을 추가로 검토해야 한다.

유사성 가정은 결과변수가 이분형일 경우 공통대조군에서의 결과발생률을 비교하여 확인할 수 있다. 만일 공통대조군에서의 결과발생률이 유사하게 나온다면 각 임상시험에서의 모집단이 유사하다고 판단할 수 있다. 공통대조군에서의 결과발생률이 다르게 나온다면 이는 연구들의 기저특성의 차이 때문일 수 있다. 이 경우 임상시험 방법론이 유사하다면 오즈비, 상대위험도 등의 상대 치료효과를 나타내는 결과 지표를 사용하여 기저특성의 차이의 영향을 줄여서 간접비교를 수행하는 것이 좋다. 결과변수가 연속형일 경우 공통대조군에서의 결과에 대한 유사성을 검토하기가 어려운데, 연속형 결과변수인 경우 변화량 등을 결과지표로 고려하는 것이 좋다.

4.3.2.3. 치료효과의 추정

공통대조군 간접비교와 혼합비교는 전통적 통계방법인 빈도론자(frequentist) 접근법 또는 베이지안 접근법으로 모두 가능하며, 이질성의 원인이 되는 공변량이 있는 경우 공변량을 보정하는 것도 가능하다.

4.3.2.4. 모형적합도 확인

치료효과 추정결과를 바탕으로 동질성, 유사성, 일관성 가정을 만족하는지 평가하고, 적용한 모형에 적절한 통계량을 이용하여 모형적합도를 확인한다. 동질성을 확인하기 위해서는 이질성을 검토하는 I^2 통계량 또는 코크란의 Q 통계량을 이용할 수 있으며, 일관성 확인을 위해서는 Bucher(1997) 등이 제안한 일관성 통계량 등을 이용할 수 있다.

4.3.2.5. 민감도 분석

치료효과에 영향을 줄수 있는 잠재적인 공변량을 추가로 고려하여 민감도 분석을 수행할 수 있으며, 특히 베이지안 접근법을 적용했을 경우 사전분포를 다르게 적용하여 민감도 분석을 실시해야 하여 결과가 강건(robust)한지 검토해야 한다.

4.3.3. 사례

2004년 AHRQ와 OMAR (Office of Medical Applications of Research)는 불면증을 가진 성인 환자에서 벤조다이아제핀(benzodiazepines, BNZ)과 비벤조다이아제핀(nonbenzodiazepines, NBNZ)의 유효성과 안정성을 비교하기 위하여 체계적 문헌고찰을 실시하고, 근거합성을 위하여 BNZ와 위약, NBNZ와 위약을 비교한 RCT들과 BNZ와 NBNZ를 비교한 RCT들을 대상으로 빈도론자 접근에 의한 직접비교, 간접비교, 혼합비교 및 베이지안 접근에 의한 직접비교, 혼합비교를 실시하였다. 이 연구에서 여러 분석방법 결과를 비교한 결과, 적용한 방법에 따라 치료효과 추정치는 유사하였으나, BNZ와 NBNZ의 치료효과의 차이는 통계적으로 유의하지 않았다. 따라서 간접비교와 혼합비교를 수행하기 위해 필요한 유사성 및 일관성 가정을 만족한다면, 직접비교와 간접비교 근거를 모두 통합하여 통합추정치의 정도(precision)를 향상시킬 수 있다.

참고문헌

김수영, 박지은, 서현주 등. 체계적 문헌 고찰 매뉴얼. 한국보건의료연구원. 2011

이현주, 권진원, 현민경 등. 비파열 뇌동맥류의 예방적 치료에 대한 성과 연구. 한국보건의료연구원. 2010

최동훈, 이상무, 서혜선 등. 급성 심근경색증 환자에서 약물방출 스텐트와 금속 스텐트의 비교. 한국보건의료연구원. 2009

한서경, 김윤희, 장은진, 안정훈, 강수희, 김선희. 간접비교를 통한 보건의료기술의 유용성 평가에서 분석결과의 타당성 확보를 위한 조건. 한국보건정보통계학회지 2011;36(2):161-171.

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683-91.

Hawkins N, Scott DA, Woods B. How far do you go? Efficient searching for indirect evidence. Med Decis Making 2009;29:273-281.

Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons & network meta-analysis for health care decision-making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value in Health 2011;14:417-428.

Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine. 2004;23:3105-24.

Song F, Loke YK, Walsh T et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: a survey of published systematic reviews. BMJ 2009:338:1-7.

Song F. What is indirect comparison? Hayward Medical Communications, a division of Hayward Group Ltd. 2009.

Vandermeer BW, Buscemi N, Liang Y, et al. Comparison of metaanalytic results of indirect, direct, and combined comparisons of drugs for chronic insomnia in adults: a case study. Med Care. 2007;45(Suppl 2): S166-S172.

PART 5

비용-효과 분석

5.1. 개요

비교효과연구에서 비용-효과 분석(cost-effectiveness analysis)은 여러 선택 가능한 의료기술들의 효과와 비용을 동시에 고려하는 방법이다. 전통적으로 보건의료 분야에서 의료기술을 평가함에 있어 효능 및 안전성이 주요 기준이 되어 왔다. 그러나 최근 고가 의료기술의 발달 및 의료비 지출의 가파른 증대로 인해 가용자원의 제약이 따르면서 효율적인 자원배분에 대한 관심이 높아졌다. 즉, 보다 효과적으로 의료기술을 사용하기 위해서는 효능 및 안전성에 관한 기준과 더불어 비용을 함께 고려한 의료기술 평가가 이루어질 필요가 있다. 단순하게 비용만 비교한 경우는 온전한 의미에서 비용-효과 분석이라고 할 수 없다. 즉 비용-효과 분석이란 효과를 고려했을 때 비용수준이 적절한지 판단하는 것으로, 비용 측면만 고려하여 가장 낮은 비용을 선택하는 방법과는 다르다고 할 수 있다.

5.2. 방법

5.2.1. 분석 단계

비용-효과 분석 단계는 크게 다음과 같이 나누어 볼 수 있다.

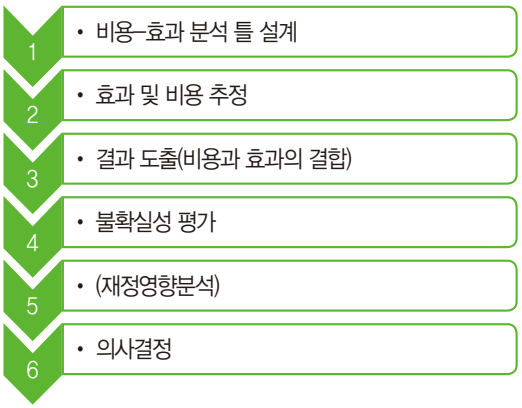

그림 5-1 비용-효과 분석 수행 단계

5.2.1.1. 비용-효과 분석 틀 설계

비용-효과 분석을 하기 위해서는 먼저 분석 목적을 분명히 하고 비교할 의료기술들을 선정하는 것이 필요하다. 비교 의료기술의 선정시 해당 의료기술이 적용될 환자군의 특성이나 질병 특성 등에 비추어 비교할 의료기술들을 적절하게 선정해야 한다. 일반적으로 현재 널리 사용되는 의료기술들을 비교하게 되나 여러 요인들도 추가로 고려할 필요가 있다. 어떤 인구집단을 대상으로 하느냐에 따라 해당 의료기술의 비용과 효과가 달라질 수 있음을 알고, 아울러 연구에서 설정한 인구집단을 기준으로 도출된 결과가 현실에서 일반화할 수 있는지 고려해야 한다. 또한 분석 관점에 따라 고려해야 하는 비용항목이 달라지므로 사전에 이를 명시해야 하며 이 외에 분석기간, 분석기법 등을 결정해야 한다. 이를 바탕으로 분석 목적에 맞는 분석 모형 틀을 설계해야 하며, 모형을 설정하는 데 있어 해당 의료기술이 적용되는 임상현실이 반영되어야 한다.

5.2.1.2. 효과 및 비용 추정

비용-효과 분석에서 고려할 수 있는 임상적 효과로는 질병의 발생, 존재와 관련된 생화학적 표지의 변화에서부터 '무증상 일수'와 같은 임상적 징후의 변화에 이르기까지 다양한 지표를 활용할 수 있다. 비용-효과 분석을 수행하는데 있어 이상적인 효과지표는 임상적으로 중요한 최종결과이다. 관련된 대부분의 외국 가이드라인(NICE, 2008; CADTH, 2006; PBAC, 2008)에서는 최종결과를 결과지표로서 선호하고 있지만, 현실적으로 최종결과에 대한 자료를 얻기 어려운 경우가 많아 대리결과 역시 결과지표의 한 형태로 받아들여지고 있다. 그러나 대리결과를 비용-효과 분석의 결과지표로 이용하는 경우, 대리결과와 최종결과간의 일관되고 강한 연관성이 입증되었거나 대리결과가 의사 결정의 기초로 간주될 수 있을 정도로 충분히 크고, 정확하고, 지속되는 것이어야 한다.

임상적 효과는 연구자가 직접 임상시험 또는 관찰연구 등을 수행하여 얻거나 이미 보고된 선행연구들을 검토하여 확인할 수도 있다. 선행연구들을 통해 효과를 추출하기 위해서 체계적 문헌고찰을 수행할 수도 있다.

비용-효과 분석에서 비용항목은 크게 의료비용, 비의료비용, 생산성 손실비용으로 나눌 수 있다. 의료비용은 의료기관 서비스를 이용하는 데 소요된 비용과 건강식품, 의료기기 구입 등 비공식적으로 지출한 비용을 모두 포함하는 개념이며, 비의료비용은 의료이용에 수반되는 비용으로 교통비용, 간병비용, 환자시간비용 등이 있

다. 생산성 손실비용은 질환 또는 질환으로 인한 조기사망으로 인해 일을 못하게 됨으로써 발생하는 경제적 손실비용을 의미한다. 2013년 한국보건의료연구원에서 발간한 「NECA 연구방법 series 4 보건의료분야에서 비용 산출방법」에서 각 비용 항목별 산출방법 및 가용 자료원에 대해 자세히 기술하고 있으므로 이를 참고한다.

미래 시점에 발생하는 비용을 현재가치화하기 위해 할인율을 적용하게 된다. 그러나 적정 할인율에 대해서는 합의된 바가 없다. 심평원의 의약품 경제성 평가지침에 따르면 기본분석에서는 5% 할인율을 적용하고, 미래 발생하게 되는 비용에 대해 할인을 하지 않는 경우, 3%, 7.5%를 적용하는 경우에 대해 별도로 민감도 분석을 수행하는 것을 제안한다. 특히 미래 발생하는 비용의 비중이 클 경우 적용하는 할인율에 따라 결과가 어떻게 달라지는 지 검토가 필요하다. 한편 해당 할인율은 물가상승률을 고려한 실질 할인율로 보고, 미래에 발생하는 비용에 대하여 임금 상승률 등 물가상승률은 추가로 적용할 필요가 없다.

5.2.1.3. 결과 도출

비용-효과 분석에서 결과 제시는 효과와 비용을 결합한 형태인 비용-효과비로 나타낸다. 비용-효과비에는 평균 비용-효과비(Average Cost-Effectiveness Ratio, ACER)와 점증적 비용-효과비(Incremental Cost-Effectiveness Ratio, ICER)가 있다. 평균 비용-효과비는 치료법 A와 치료법 B의 두 가지 대안이 있을 때, A 대안의 평균 비용-효과비와 B 대안의 평균 비용-효과비를 각각 구하여 비교하는 것이다. 이 경우 비용-효과비가 작은 대안일수록 비용-효과적이라고 할 수 있다. 반면 점증적 비용-효과비의 경우 치료법 A와 치료법 B를 비교하여 효과의 차이를 비용의 차이로 나눈 값이다. 즉, 이 경우 점증적 비용-효과비는 효과 1단위 개선하는데 얼마만큼의 비용이 추가로 필요한지 나타내는 지표이다. 평균 비용-효과비는 두 가지 대안의 비를 산출하여 비교함으로써 비용-효과성을 판단하는 반면, 점증적 비용-효과비는 각각 산출된 효과와 비용의 차이로 하나의 비를 산출한 후 사회적으로 수용 가능한 임계값과의 비교를 통해 비용-효과성을 판단하게 된다. 일반적으로 대안들이 상호경쟁 관계에 있는 대안들에 대하여 비용-효과 분석이 이루어지므로 점증적-비용 효과비를 이용한다.

일반적으로 점증적 비용-효과비를 가장 많이 활용하지만 비용이 떨어지면서도($\triangle C<0$) 효과가 더 개선되는 상황($\triangle E>0$)이나 비용이 증가하면서도($\triangle C>0$) 효과가 떨어지는 상황($\triangle E<0$)에서는 점증적 비용-효과비를 산출해서는 안 된다. 비용이 △-600만원이면서 효과가 △2년인 경우 단순하게 비를 산출하면 -300만원이

되며 비용이 △600만원이면서 효과가 △-2년인 경우에도 비는 -300만원이 된다. 그러나 전자의 경우 비용이 더 낮아지면서 효과가 더 개선되는 상황으로 어느 면으로 보나 우월한 대안이 되며 후자의 경우 열등한 대안이 되기 때문에 점증적 비용-효과비가 비용-효과성을 판단하는 적절한 기준이 되지 못한다. 따라서 점증적 비용-효과비를 이용한 비용-효과성 판단은 비용-효과성 판단이 필요한 상황(효과도 증가하고 비용도 증가하거나 효과는 감소하나 비용도 감소하는 상황)에서만 가능하다.

점증적 비용-효과비 외에 비용-효과 분석 결과를 제시할 때 사용할 수 있는 방법으로 비용-효과 수용 곡선(cost-effectiveness acceptability curve)과 순건강편익(net health benefit)이 있다.

5.2.1.4. 불확실성 평가

비용-효과 분석은 많은 연구들과 가정의 복잡한 결합이므로 불확실성을 고려해야 한다. 민감도 분석(sensitivity analysis)이란 비용 또는 효과추정에 사용된 변수의 불확실성을 확인하기 위한 대표적인 방법이다. 민감도 분석은 포함된 변수 값의 변화가 결과에 미치는 영향을 확인함으로써 의사 결정자에게 전반적인 불확실성의 정도 뿐 아니라 결과에 크게 영향을 미치는 변수에 대한 정보를 제공한다. 하나 또는 몇 개의 변수를 변화시키면서 결과의 변화를 확인하는 단순 민감도 분석과 변수들의 분포를 가정하여 주어진 확률 분포로부터 무작위로 값을 추출하여 이를 바탕으로 변화를 확인하는 확률적 민감도 분석이 있다.

5.2.1.5. 재정영향 분석

비용-효과 분석과 더불어 필요한 경우 재정영향분석을 선택적으로 수행할 수 있다. 재정영향분석이란, 의료기술의 급여 여부 변화에 따라 건강보험 재정에 미치는 영향을 추정하는 방법으로 보험자와 정책 결정자에게 예산 관리에 있어 의미 있는 정보를 제공한다. 재정영향 분석 수행시 비교하는 의료기술들이 적용되는 환자집단과 비교 의료 의료기술들의 가격 등을 고려하게 된다. 일반적으로 급여 변화 후, 3년~5년간의 건강보험재정에 미치는 영향을 분석하게 되며 단순하게 해당 의료기술들의 비용과 더불어 효과 개선 등으로 치료비용 등이 감소했다면 이를 반영하여 추정하는 것이 더 바람직할 것이다.

5.2.1.6. 의사 결정

보건의료 분야에서 비용-효과 분석 결과는 주로 서로 경쟁적인 대안들간 비용-효과성을 확인하는 지표인 점증적 비용-효과비로 산출된다. 이렇게 산출된 점증적 비용-효과비를 토대로 이 의료기술이 실제 수용 가능한 것인지 판단하기 위해서는 의사 결정과정이 필요하다. 수용 가능한 비용-효과비에 대해서는 각국의 의료 환경과 경제 수준이 다르므로 전 세계적으로 통용되는 보편적인 기준이 없고, 다만 각 국가별로 급여결정에 있어 미국은 $50,000, 영국은 £20,000~30,000와 같이 어느 정도가 기준점일 것이라는 암묵적인 합의들이 있다. 우리나라의 경우 비용-효과 분석이 점차 널리 확대되고 있으며 1인당 GDP 수준을 암묵적으로 사용하고 있다.

그러나 의료기술 수용여부에 관한 의사결정과정에서 비용-효과 분석 결과만이 유일한 기준이 되어서는 안 된다(Clark and Weale, 2012). 비용-효과 분석 결과는 객관적으로 제시할 수 있는 합리적인 근거 중 하나이면서 의사결정을 하는 데 있어 여러 고려사항 중 하나임을 명심해야 한다.

5.2.2. 비용-효과 분석 유형

보건의료에서 경제성 분석 방법으로 비용최소화 분석, 비용-효과 분석, 비용-효용 분석, 비용 편익 분석이 있다. 비용-효과 분석이란 용어가 미국에서는 이를 모두 포함하는 개념의 넓은 의미로 사용되는 반면 유럽에서는 주로 좁은 의미로서 경제성 분석 방법의 하나로 언급된다.

비용-최소화 분석(cost-minimization analysis)은 효과를 고려했을 때 동일한 수준이라고 인정되면 비용만 비교하는 방법이다. 이 경우 비용 분석에 앞서 효과가 동일한 수준인지 판단하는 과정이 제시되어야 한다. 효과가 동일하다고 판단되면 비용이 가장 저렴한 대안이 가장 효율적인 대안이라고 할 수 있다. 일반적으로 성분이 동일한 제네릭 약물들의 경우 치료 약제들간 결과가 동일하다고 볼 수 있으므로 비용-최소화 분석을 적용할 수 있다. 때로는 유사 개량약의 비교에도 사용할 수 있다. 유사개량약이라 함은 분자구조에는 약간 차이가 있으나 작용기전이 동일하고 임상적 효과도 별도 다르지 않는 약이다. 한편 임상시험에서의 효능이 동일한 것으로 보고되었더라도 순응도에 영향을 미치는 투여 방법이나 투여 횟수 등의 차이로 실제 임상 현장에서 효과가 차이가 난다면 다른 분석 방법을 고려해 볼 수 있다. 또한 비열등성 임상시험 설계에 기초하여 효과가 동일한지 판단할 때 통계적 검정력이 충분하지 않아 차이를 발견하지 못한 경우 이를 비용-최소화 분석으로 수행하게

된다면 잘못된 의사 결정으로 이끌 수 있다. 이런 상황에서는 신뢰구간을 고려한 비용-효과 분석이 더 적절할 수 있다.

비용-효과 분석이나 비용-효용 분석(cost-utility analysis)은 효과의 크기가 다를 때 효과의 차이와 비용의 차이를 함께 고려하여 경제성을 평가하는 방법이다. 비용-편익 분석(cost-benefit analysis)과 다른 점은 결과가 건강 결과 그 자체로 평가된다는 점이다. 비용-효과 분석은 비용을 자연단위(natural unit)로 측정된 효과와 비교하는 방법이다. 효과의 크기는 다르더라도 측정 단위는 동일해야 비교가 가능하므로 효과의 단위가 동일한 상황인 경우만 비교가 제한된다는 단점이 있다. 비용-효과 분석에 사용하는 단일한 결과 측정치는 혈압감소율, 혈청콜레스테롤 감소율과 같은 생리학적 지표의 변화에서부터 무증상일, 발병률 감소, 사망률 감소 등 최종 결과에 이르기까지 다양하다.

비용-효용 분석은 질보정수명(Quality Adjusted Life Years, QALY), 건강수명(Healthy Years Equivalents, HYE)과 같이 삶의 양적인 측면(수명연장)과 질적인 측면(삶의 질)을 함께 반영한 지표를 사용한다. 비용-효과분석에 비해 삶의 질 수준을 수명 연장과 함께 고려할 수 있다는 점에서 삶의 질 개선이 주요 효과인 치료법의 경우 유용한 분석 방법이 된다.

표 5-1 비용-효과 분석 방법의 종류

분석 유형	비용측정가치평가	결과의 종류	결과의 측정가치 평가
비용-최소화 분석	화폐단위	모든 면에서 동일 (효과까지 동일)	같다는 것을 보여줌
비용-효과 분석	화폐단위	대안들에 공통된 단일 효과 적용 그러나 효과 개선 정도는 다름	수명연장, 혈압 감소
비용-효용 분석	화폐단위	단일 효과 혹은 다중 효과 적용 그러나 효과 개선 정도는 다름	질보정생존년수 건강수명
비용-편익 분석	화폐단위	단일 효과 혹은 다중 효과 적용 그러나 효과 개선 정도는 다름	화폐단위

(자료원 : Drummond 등, 2005. 재편집)

비용-편익 분석은 건강수준의 개선, 생존기간의 연장 같은 편익을 건강 그 자체를 나타내는 지표가 아닌 화폐단위로 계량화하여 평가하고 비교한다. 이론적으로 서로 다른 분야간 프로그램의 비교도 가능한데 이는 결과가 모두 화폐단위로 표현되기 때문이다. 또한 비용과 효과를 모두 화폐단위로 표현하므로 그 자체로 결과를 판단할 수 있다. 즉 어떤 프로그램에서 비용보다 화폐단위로 추정한 편익이 더 크다면 그 자체로 경제성이 있다고 판단할 수 있다. 비용-편익 분석은 교통이나 환경 분

야에서 특정 프로그램에 대한 경제성을 판단할 때 많이 사용하지만 보건의료분야에서는 자주 사용되지 않는 기법이다. 보건의료분야에서 인간 생명에 화폐적 가치를 부여하는 것에 대하여 여러 이슈가 존재하며 화폐적 가치를 추정하기 위한 방법론에 대해서도 윤리적 문제가 작용하기 때문이다.

5.2.3. 비용-효과 분석 모형

효과와 비용 등 여러 정보들을 종합, 합성하여 비용 대비 효과를 확인하는 것은 주로 모형 구축을 통해 이루어진다. 모형은 임상시험에서 관찰된 대리지표의 변화를 최종결과의 변화로 연동하거나 관찰기간을 넘어서 장기적 결과를 예측하고자 할 때 사용될 수 있다. 또한 통제된 환경에서 측정한 효능으로부터 실제진료환경에서 기대할 수 있는 효과를 추정하는 과정이나 다른 환경에서 분석된 결과를 의사결정이 이루어지는 환경에 적합하게 하는 과정에서도 모형을 이용한다(박병주 등, 2009).

비용-효과 분석에서 결정분석 모형(decision analytic model)을 주로 이용하는데 이것은 불확실한 상황에서 이루어지는 의사 결정을 위한 체계적 접근법으로, 하나 혹은 그 이상의 서로 다른 대안들이 가지는 상대적 가치를 분석하는 방법이다. 결정분석 모형에서 각 대안에 따른 결과(건강상태)가 일어날 가능성을 확률로 표현한 후, 이들 결과에 따른 비용과 삶의 질 등의 정보를 결합하여 각 대안의 기대비용과 기대결과를 계산하게 된다. 결정분석 모형에서는 결정수형 모형(decision tree)과 마콥 모형(Markov model)이 흔히 이용된다.

5.2.3.1. 결정수형 모형

결정수형 모형은 불확실한 상황 하에서 의사결정자가 내리는 선택과 그로 인해 발생 가능한 결과들을 그림으로 표현한 것이다. 관련대안과 각 대안들을 선택했을 때 이어지는 일련의 확률적 사건들, 그러한 사건이 발생할 확률과 최종결과로 구성된다.

결정수형 모형은 네 가지 기본 요소로 이루어진다. 결정마디(□)는 의사결정자가 하나 혹은 그 이상의 가능한 경로에 대한 선택을 하게 되는 지점이며 확률마디(○)는 의사 결정자의 통제를 넘어선 불확실한 결과를 표현할 때 사용된다. 각 확률마디마다 사건들이 일어날 가능성인 확률이 부여되며 결정수형의 가지 끝에 각 경로의

결과를 할당하여 분석하게 된다.

그러나 사건이 반복되고 예후가 복잡한 경우(만성질환의 경우) 단순한 결정수형만으로 분석이 어려울 수 있다. 또한 고려해야 하는 건강상태가 많아 모든 가능한 경우의 수를 고려해야 할 때 모형구조가 너무 복합해지는 문제점이 있다. 시간단위가 분명히 정의되지 않아 할인을 하거나 QALY처럼 생존기간을 보정하는 것에 어려움이 있다.

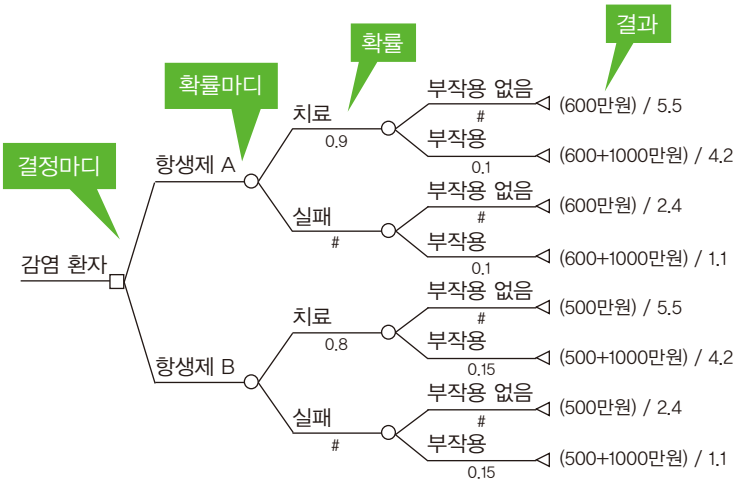

그림 5-2 결정수형 모형과 구성 예시

5.2.3.2. 마콥 모형

마콥 모형에서 환자는 시간의 흐름에 따라 일정한 확률로 서로 다른 건강상태로 이동하고 각 상태에 머무는 동안 발생하는 비용과 삶의 질 등이 시간의 경과에 따라 누적된다. 마콥 모형 구축시 건강 상태(마콥 상태), 분석주기, 전이확률을 결정해야 한다.

건강 상태(마콥 상태)를 정의함에 있어 질병 또는 치료과정과 관련된 주요 건강상태를 포함해야 하며 각 상태는 서로 배타적이어야 한다. 분석주기는 각 건강상태(마콥 상태)에서 다른 건강상태(마콥 상태)로의 이전이 일어나는 고정된 시간 간격으로 질병의 특성에 따라 달라진다. 예를 들어 질병의 경과가 빠르게 변하는 경우 주기가 짧으며 만성질환처럼 변화속도가 느린 경우 긴 주기(보통 1년)를 선택하게 된다. 전이확률은 환자가 특정 건강상태(마콥 상태)에서 다음 주기에 다른 건강상태(마콥 상

태)로 이동할 확률로 각각의 건강상태로 이동하는 확률의 합은 1이 되어야 한다. 전이확률은 환자가 이전에 어떤 경험을 하였는지 상관없이 현재 어떤 상태에 있는지에 따라 다음 단계로 이동할 확률이 결정된다(박병주 등, 2009).

모형 분석은 크게 일정한 특성을 가진 인구집단을 대상으로 하는 코호트 시뮬레이션(cohort simulation) 방법과 한 사람 한 사람을 전이확률에 따라 모형에 포함된 각 건강상태로 시뮬레이션하는 마이크로 시뮬레이션(micro simulation) 혹은 몬테카를로 시뮬레이션(Monte Carlo simulation) 방법을 통해 이루어진다. 시뮬레이션의 수가 많아지면 코호트 시뮬레이션 결과와 마이크로 시뮬레이션 한 결과에서의 점 추정치는 거의 일치하게 된다.

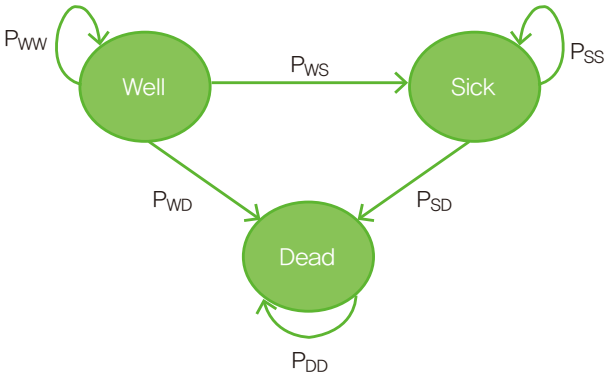

그림 5-3 마콥 모형
(자료원: 박병주 등(2009) 재편집)

5.2.4. 비용-효과 분석 연구 평가

경제성 평가 연구에 대한 비판적 평가를 수행하기 위한 몇 가지 도구들이 보고된 바 있다(NHS, 2002; Drummond 등, 2005; Evers 등, 2005). 본 연구에서는 영국 NHS의 'Critical Appraisal Toolkit for Economic Evaluation'을 일부 수정하여 사용한 안정훈 등(2010)의 도구를 제안한다.

표 8-2 비용-효과 분석 문헌의 비판적 평가 문항

A. Is the economic evaluation likely to be usable?
1. Was a well-defined question posed? Is it clear what the authors are trying to achieve?
2. Was a comprehensive description of the competing alternatives given? Can you tell who did what to whom, where and how often?
3. Effectiveness a) Does the paper provide evidence that the programme would be effective (i.e. would the programme do more harm than good)? Consider if a RCT was used; if not, consider how strong the evidence was. b) Was the effectiveness measured and valued appropriately?
B. How were consequences and costs assessed and compared?
4. Were all important and relevant resource use and health outcome consequences for each alternative: a) Identified? Consider what perspective(s) was / were taken b) Measured accurately in appropriate units prior to evaluation? Appropriate units may be hours of nursing time, number of physician visits, years-of-life gained etc. c) Valued credibly? Have opportunity costs been considered?
5. Were resource use and health outcomes consequences adjusted for different times at which they occurred (discounting)?
6. Was an incremental analysis of the consequences and costs of alternatives performed?
7. Was an adequate sensitivity analysis performed? Consider if all the main areas of uncertainty were considered
C. Will the results help in purchasing for local people?
8. Did the presentation and discussion of the results include enough of the issues that are required to inform a purchasing decision?
9. Were the conclusions of the evaluation justified by the evidence presented?
10. Can the results be applied to the local population? Consider whether: - The patients covered by the review could be sufficiently different to your population to cause concern - Your local setting is likely to differ much from that of the review
11. Was the model validated? (clinical relevance/logic, coding, extrapolation techniques, transparency of path ways)

(자료원: 안정훈 등(2010), 영국 NHS에 3b)와 11의 항목을 추가 재편집)

5.3. 사례

본 장에서는 한국보건의료연구원에서 수행된 연구 중 급성 심근경색증 환자에서 약물방출 스텐트와 금속 스텐트를 비교한 연구(최동훈 등, 2009)와 치료 경로를 고려한 항우울제군간 비용-효과 분석을 수행한 연구(안정훈 등, 2010)를 소개하고자 한다.

최동훈 등(2009)의 연구에서는 급성 심근경색증 환자에서 약물방출 스텐트와 금속 스텐트의 임상적 효과에 대한 체계적 문헌고찰을 통해서 주요 결과변수인 사망에 있어 두 스텐트간 유의한 차이가 없는 것으로 확인하였다. 이를 근거로 비용-효과 분석방법 중 비용 최소화 분석을 수행하였다. 사회적 관점으로 분석을 수행하였으며 결정 수형 모형을 활용하였고, 분석 기간은 1년으로 설정하였다. 모형 분석에 필요한 전이확률 등은 건강보험심사평가원 청구자료 분석 결과를 적용하였다. 비용 최소화 분석 결과 두 스텐트 간 재시술률의 차이 등으로 인한 비용 등을 모두 고려하였을 때, 금속 스텐트 시술시 소요비용이 약물 방출 스텐트로 시술하는 경우보다 낮은 것으로 나타났다. 그러나 이러한 결과는 재시술을 받지 않은 환자의 비용과 비율에 민감하게 달라져 추가 연구가 필요한 것으로 결론을 내리고 있다.

안정훈 등(2010)은 치료경로를 고려하여 항우울증제인 삼환계 항우울제(Tricyclic Antidepressants, TCA)와 선택적 세로토닌 재흡수 억제제(Selective Serotonin Reuptake Inhibitors, SSRI), 새로운 항우울제(New Anti-Depressants, NAD) 군간의 비용-효과 분석을 수행하였다. 약물군 간의 임상적 효과 및 의료비용 산출을 위해서, 우울증 신환자를 대상으로 약물치료경로에 따라 3개월의 약물 투약기간 동안 항우울제 복용에 순응여부와 순응여부에 따른 재발률 등을 건강보험심사평가원 청구자료를 이용하여 분석하였다. 사회적 관점 및 보험자 관점에서 각각 분석을 수행하였으며 분석모형은 결정 수형 모형을 활용하였다. 3개월 동안 약물 순응여부를 확인하고 약물 복용기간이 종료된 3개월 이후 재발을 파악하는 기간을 9개월로 하여 전체 분석기간은 1년으로 하였다. 비용-효과 분석 결과 우울증 치료시 항우울제 중 SSRI계열의 약물로 치료를 시작하는 것이 우리나라의 비용-효과성 판단 기준에 근거할 때 가장 비용-효과적인 치료대안인 것으로 나타났으며, 이러한 결과는 민감도 분석에서도 일관되게 나타났다. 그러나 심한 불안, 불면, 신체증상을 동반하는 등 임상적으로 항우울제 선택이 필요한 환자는 본 비용-효과 분석에서 설정한 환자 대상군과 다르므로 본 연구 결과를 적용하는데 주의가 필요하다고 언급하고 있다.

참고문헌

건강보험심사평가원. 의약품 경제성평가지침 및 자료작성요령. 2011.

박병주, 허대석, 안형식, 이상무, 윤영호, 김수영 등. 근거중심 보건의료. 고려의학. 2009.

안정훈, 김윤희, 신상진, 박선영, 송현진, 박주연, 배은영. 근거중심의 진료에 맞는 한국적인 보건의사결정을 위한 방법론 연구. 한국보건의료연구원. 2010.

안정훈, 정선영, 신상진, 송현진, 박주연, 백종우, 서국희, 김주영, 한창수, 우종민, 이민수, 김남순, 김경미, 김철민, 정승희. 항우울제 투약순응도에 영향을 미치는 요인분석 및 경제성 평가. 한국보건의료연구원. 2011.

최동훈, 이상무, 서혜선, 송현진, 장은진, 최지은, 손희정, 김중선, 안영근, 정명호, 황혜진. 급성 심근경색증 환자에서 약물방출 스텐트와 금속 스텐트의 비교. 한국보건의료연구원. 2009.

Clark S, Weale A. Social values in health priority setting: a conceptual framework", Journal of Health Organization and Management. 2012;26(3): 293–316.

Drummond MF, Schulper M, Torrance GW, O'Brien B, Stoddart GL. Methods for the economic evaluation of health care programmes: Oxford University Press. 2005.

Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. Int J Technol Assess Health Care. 2005;21(2):240–245.

Gold MR, Patrick DL, Torrance GW, Fryback DG, Hadorn DC, Kamlet MS, Daniels N, Weinstein MC, Identifying and valuing outcomes. cost–effectiveness in Health and Medicine. New York(NY): Oxford University Press; 1996:82–134.

Guide to the methods of technology appraisal. UK National Institute for Health and Clinical Excellence. June 2008.

Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.3). Commonwealth of Australia Pharmaceutical Benefits Advisory Committee; December 2008

Guidelines for the economic evaluation of health technologies: Canada [3rd Edition]. Ottawa: Canadian Agency for Drugs and Technologies in Health. 2006.

National Health Service. Critical Appraisal Toolkit for Economic Evaluation. 2002.

Ofman JJ, Sullivan SD, Neumann PJ, Chiou CF, Henning JM, Wade SW, Hay JW. Examining the value and quality of health economic analyses: implications of utilizing the QHES. J Manag Care Pharm. 2003;9(1):53–61.

PART 6

특정 분야의 비교효과연구

6.1. 진단 정확도에 대한 비교효과연구

6.1.1. 개요

새롭게 개발된 진단검사가 기존 진단검사와 비교하여 어떤 상황에서 어떤 환자의 임상성과를 향상시킬 수 있을지에 대한 근거에 대한 요구가 많아지고 있다. 하지만, 진단검사에 대한 무작위배정임상연구는 부족한 상황이다.

진단검사는 환자의 임상적 상황, 다른 진단검사의 결과, 진단검사를 시행하는 시술자의 지식과 기술, 치료법의 적용가용성 등 복합적인 상황에서 사용되고 있다. 이런 복합적인 상황에서 비교효과연구를 수행하는 연구자들은 기술적인 사양에 따른 진단검사의 능력, 진단 혹은 예후 분류기준에 따라 환자를 정확하게 나눌 수 있는 능력, 임상가와 환자의 생각이나 행동에 영향을 줄 수 있는 것 등 매우 좁고 세밀한 문제를 고려하여 관련 근거를 광범위하게 모두 수집해서 비교 평가해야 한다.

6.1.2. 방법

여기에서는 진단법 체계적 문헌고찰에 대해 서술하기로 한다.

6.1.2.1. 진단법 체계적 문헌고찰

기존의 근거를 수집하여 진단법을 평가하는 방법으로 진단법 체계적 문헌고찰이 있다. 진단법 체계적 문헌고찰의 목적은 건강성과에 대한 진단검사의 영향을 평가하기 위하여 근거를 규명하고 통합하여 의사결정을 돕는 것이다. 이를 위해 먼저 검토하고자 하는 검사법에 대해 분명한 기술이 필요하다. 예를 들어 건강한 사람들에 대한 검진으로서의 검사인지 진단을 위한 검사인지 분명히 기술해야 하며, 검사의 역할 즉 생검 실시여부와 같은 의사결정에 대한 검사법의 역할 등에 대해서도 고려할 필요가 있다. 두 번째로 진단법은 임상성과에 간접적으로 영향을 미치므로 어떤 간접적인 결과가 임상성과에 영향을 미치는지 파악하는 것이 중요하다. 비교하고자 하는 검사법을 환자에게 무작위배정을 실시하여 시행한 무작위배정 비교임상시험에서 최종 건강성과까지 평가하여 진단법의 직접효과를 파악하는 것은 거의 힘들기 때문에, 일반적으로 검사정확도와 같은 중간성과로 평가한다.

위에서 언급한 두 가지를 다루기 위한 원칙은 다음과 같다.

첫째, 연구목적을 분명히 하고, 연구범위와 목적에 맞는 연구질문을 결정한다. 이때 가능한 자원과 시간을 고려하여 주제의 중요성과 실현가능성의 균형을 맞춰야 한다.

둘째, 분석틀(analytic framework) 또는 인과경로(causal pathway)를 개발한다. 분석틀에서 연구대상, 중재, 임상성과 등을 그림으로 나타냄으로써 직접적인 근거가 부족할 때 잠재적인 연구가설까지 표현할 수 있으므로 유용하다(그림 6-1).

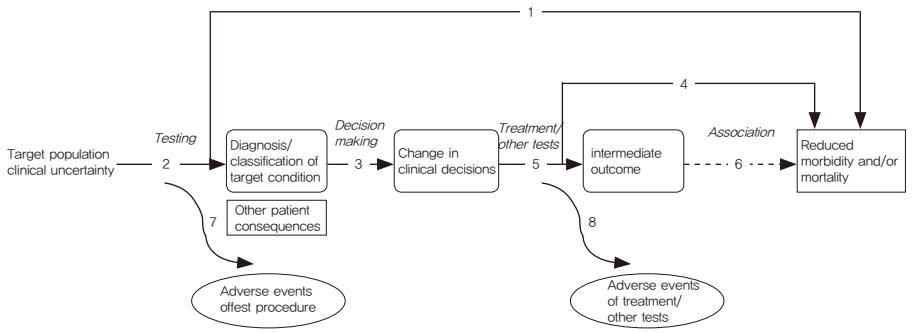

그림 6-1 USPSTF에서 제시한 진단검사평가에서의 분석틀(Chang, 2012)

Reprinted from Current methods of the U.S. Preventive Services Task Force: A review of the process, 20(3), Russell P. Harris,Mark Helfand,Steven H. Woolf,Kathleen N. Lohr,Cynthia D. Mulrow,Steven M. Teutsch,David Atkins, 21–35, Copyright (2013), with permission from Elsevier. Adapted from Chang SM, Matchar DB, Smetana GW, Umscheid CA, editors. Methods Guide for Medical Test Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012.

셋째, 결정수형모형(decision tree)을 사용하여 관심 진단법 이외 비교검사법을 고려한다. 결정수형모형은 진단정확도 지표와 가능한 경로와 성과를 규명함으로써 핵심연구질문을 분명하게 하는데 도움을 준다.

넷째, 정확도 이외 성과까지 연구질문에 포함하지 않고, 정확도 연구만 고려하여 평가해도 충분한 경우가 있다. 이때 고려해야 하는 사항은 다음과 같다.

표 6-1 정확도 관련 고려사항

항목	고려사항
1	새로운 검사에서 기존 검사보다 추가적인 환자를 진단을 할 수 있는가?
2	연구에서 선정된 환자들에게 새로운 검사법을 적용할 수 있는가?
3	새로운 검사가 반응을 예측하는지를 연구에서 평가할 수 있는가?
4	가용할 수 있는 연구가 기존 검사로 평가한 환자뿐일 경우, 새로운 검사로 추가 진단된 환자는 연구 참여자들과 동일한 스펙트럼이나 질병분류를 나타내는가?
5	검사결과는 참고표준(reference standard)검사에 의해 확인되었나?
6	새로운 검사는 질병의 정의나 스펙트럼(예. 조기진단)을 바꿀 수 있나?
7	새로운 검사와 기존 검사의 정확성과 치료효과가 이질적인가?

진단법 정확도 체계적 문헌고찰(systematic reviews of diagnostic test accuracy)은 개별 진단법 정확도 관련 연구를 체계적으로 합성하여 해당 진단법의 정확도를 평가하는 방법이다. 진단법 정확도 체계적 문헌고찰은 일반적인 중재법의 체계적 문헌고찰 방법을 그대로 따르나 진단법 정확도 연구에서 특별히 고려해야할 사항은 다음과 같다.

(1) 선정기준

연구하고자 하는 연구질문에 대해 다음과 같이 명확히 하도록 한다.

- 연구대상자(Participants) : 평가하고자 하는 진단법을 적용하는 연구대상자
- 시험 진단법(Index test) : 정확도를 평가하고자 하는 진단법
- 비교 진단법(Comparator test) : 일반적으로 실제 임상에서 진단 시 시험 진단법과 비교해서 사용할 만한 진단법. 필수는 아님
- 진단하고자 하는 질환 혹은 상태(Target condition) : 시험 진단법을 통해서 진단하고자 하는 질환 혹은 건강 상태
- 참고 표준(reference standard) : 이른바 '황금 표준(gold standard)'으로 검사한 대상자들에게 진단하고자 하는 질환이 있는지 없는지 확정해주는 진단법
- 연구 형태(types of studies) : 체계적 문헌고찰에서 포함되는 연구 형태. 예를 들어 환자-대조군 연구. 모든 형태의 연구

(2) 검색

진단법 정확도 연구는 다른 목적의 연구 내에 숨겨져 있는 경우가 많기 때문에 일반적으로 중재법 체계적 문헌고찰에 비해 폭 넓은 검색을 수행해야한다. 따라서 민감도를 최대한 높여서 검색하도록 하는데 일반적으로 시험 진단법(index test)과 진

단하고자 하는 질환 혹은 상태(target condition)에 대해 검색하도록 한다.

(3) 질 평가 (비뚤림 위험 평가)

선정된 진단법 정확도의 문헌에 대한 질평가 혹은 비뚤림 위험 평가는 QUADAS II를 이용하여 평가한다(Whiting 등, 2011).

(4) 자료의 합성

진단법 정확도는 민감도(sensitivity)와 특이도(specificity)로 표현되는데 이를 합성하는 메타분석은 '6.1.2.3. 진단법 정확도의 메타분석'을 참고하도록 한다.

(5) 근거의 수준 및 권고사항

중재법의 경우 근거의 수준을 도출하기 위해 일반적으로 권고 등급 도구인 GRADE(The Grading of Recommendation, Assessment, Development and Evaluation)를 많이 사용한다. 진단법 정확도 체계적 문헌고찰에서는 아직 GRADE가 확정되지 않았다. 다만 검진의 경우 미국 USPSTF에서 사용하는 검진 권고 등급을 사용할 수 있으며, 진단법 정확도의 경우 이를 참고할 수 있다.

자세한 사항은 2014년에 발간될 「NECA 진단법 정확도 체계적 문헌고찰 매뉴얼(가칭)」을 참고하도록 한다.

6.1.2.2. 진단법 정확도 분석

일반적으로 진단법 정확도의 결과는 〈표 6-2〉과 같이 진양성(true positive, TP), 위양성(false positive, FP), 진음성(true negative, TN), 위음성(false negative, FN)로 나타내며, 이를 바탕으로 진단법 정확도 평가를 위해 〈표 6-3〉에서 정의된 민감도, 특이도, 양성예측도, 음성예측도, 양성 우도비, 음성 우도비를 평가지표로 고려한다.

표 6-2 진단검사 결과와 질환의 이원분류표

		질환(reference standard)		합계
		유(D+)	무(D-)	
진단검사결과	양성(T+)	진양성(TP)	위양성(FP)	TP+FP
	음성(T-)	위음성(FN)	진음성(TN)	FN+TN
합계		TP+FN	FP+TN	N

표 6-3 정확도 지표

성과지표	산출식	정의
민감도	TP/(TP+FN)	질환이 있는 사람 중 검사결과가 양성인 사람의 비율
특이도	TN/(FP+TN)	질환이 없는 사람 중 검사결과가 음성인 사람의 비율
양성예측도	TP/(TP+FP)	검사 결과가 양성인 사람 중 실제 질환에 걸린 사람의 비율
음성예측도	TN/(FN+TN)	검사 결과가 음성인 사람 중 실제 질환에 걸리지 않은 사람의 비율
양성 우도비 (양성결과의 우도비)	민감도/(1-특이도)	질환을 가지지 않은 사람이 이상소견을 보일 확률에 비해 질환을 가진 사람이 이상소견을 보일 확률
음성 우도비 (음성결과의 우도비)	(1-민감도)/특이도	질환을 가지지 않은 사람이 이상소견을 보이지 않을 확률에 비해 질환을 가진 사람이 이상소견을 보이지 않을 확률

진단 오즈비(diagnostic odds ratio, DOR)는 진단법 정확도를 하나의 값으로 요약하는 지표로 질환이 있는 사람 중 결과가 양성이 될 오즈(odds)가 질환을 가지지 않은 사람 중 결과가 양성이 될 오즈에 비해 몇 배 높은지를 나타내는 오즈비이다. 진단 오즈비는 오즈비로 표현되기 때문에 임상적인 해석이 직관적이지 않으므로 진단법 정확도에 대한 개별 연구에서는 거의 사용되지 않으나, 진단법 정확도 메타분석 방법에서는 중요하게 사용된다. 진단검사 결과의 양성, 음성을 정의하는 역치(threshold value)에 따라 민감도와 특이도는 서로 의존적으로 변하는데, 민감도가 증가하면 특이도가 감소하고 특이도가 증가하면 민감도가 감소한다. 따라서 ROC(receiver-operating characteristic) 곡선을 사용하여 역치의 변화에 따른 민감도와 특이도를 나타내며, 곡선하 면적(area under the curve, AUC)으로 정확도를 요약할 수 있으며 AUC가 1에 가까울수록 정확도가 높은 검사라고 할 수 있다.

6.1.2.3. 진단법 정확도의 메타분석

진단법 정확도에 대한 메타분석은 일반 중재법의 메타분석과 달리 정확도를 나타내는 민감도와 특이도를 동시에 고려하여야 하므로, 먼저 개별 연구들의 민감도, 특이도를 계산하고 forest plot을 작성한다. 또한 이를 바탕으로 SROC(summary ROC) 곡선에 개별 연구들의 정확도를 표시하는데, 이때 개별연구의 각 점(point)의 크기는 정도(precision)를 나타내며 가로의 길이는 특이도의 정도, 세로의 길이는 민감도의 정도를 나타낸다. 추정치의 표준오차의 역수를 정도를 나타내는 척도로 사용하는 경우, 각 점의 가로, 세로 길이는 각 추정치의 표준오차의 역수라고 할 수 있다. Moses 등(1993)이 제안한 SROC 곡선은 진단 오즈비를 종속변수로, 전체 양성결과의 비율(연구별 역치의 차이를 반영하는 대리변수)을 독립변수로 하는 단

순선형회귀모형에서 추정된 회귀계수를 이용하여 민감도의 기대값을 계산하는 방법으로, 가중치로 로그(진단오즈비)의 분산의 역수를 사용한다. SROC 곡선은 민감도와 특이도의 통합 추정치는 제공하지 못한다는 제한점이 있지만, RevMan를 사용하여 비교적 쉽게 개별 연구들의 민감도와 특이도를 탐색적으로 분석할 수 있다.

진단법 메타분석에서 통합 추정치를 추정하고자 할 경우, 연구간 이질성을 반영하기 위해 랜덤효과모형을 사용하기를 권고하는데(Macaskill 등, 2010), 연구간 이질성 및 민감도와 특이도의 상관성까지 반영하기 위하여 Reitsma 등(2005)이 제안한 이변량 랜덤효과모형(bivariate random effect model)을 사용하여 통합 추정치를 추정한다. 이변량 랜덤효과모형에서 추정한 logit(민감도), logit(특이도), logit(민감도)의 분산, logit(특이도)의 분산, logit(민감도)와 logit(특이도)의 공분산을 RevMan에 입력하여 통합 추정치를 SROC 곡선에 표시하며, logit(민감도)의 표준오차, logit(특이도)의 표준오차, 이들의 공분산의 표준오차를 추가적으로 RevMan에 입력하여 95% 신뢰영역(confidence region)과 95% 예측영역(prediction region)을 SROC 곡선에 나타낸다. 이때 통합 민감도와 통합 특이도는 평균 정확도를 나타내며, 예측영역은 모형이 맞을 경우 미래 연구에서의 민감도와 특이도가 위치할 영역으로 통계적 이질성을 반영한다.

진단법의 정확도는 역치에 영향을 받고, 통합 대상 연구들의 역치가 이질적인 경우, 역치의 변화에 따른 ROC 곡선을 통합하기 위하여 Rutter(1995, 2001)가 제안한 HSROC (hierarchical summary ROC) 곡선 모형을 사용할 수 있다.

6.1.3. 사례

한국보건의료연구원은 2012년 류마티스 관절염 환자에서 항 CCP 항체(anti-cyclic citrullinated protein antibodies, anti-CCP Ab) 검사의 임상적 유용성을 평가하는 연구를 진행하였다. 본 연구에서는 한국인을 대상으로 한 항 CCP 항체 검사의 진단 정확도에 대한 임상적 효과 평가를 위해 기존의 체계적 문헌고찰을 활용한 문헌고찰을 수행하였다. 전체대조군을 대상으로 한 메타분석 결과, 통합 민감도는 0.76(95% 신뢰구간: 0.73-0.79), 통합 특이도는 0.96(95% 신뢰구간: 0.93-0.97), 통합된 민감도와 특이도로 계산된 양성 우도비는 18.04(95% 신뢰구간: 11.80-27.57), 음성 우도비는 0.25(95% 신뢰구간: 0.23-0.28)로 나타났다(현민경 등, 2012).

6.2. 감염질환에 대한 비교효과연구[4)]

6.2.1. 개요

항생제내성균에 의한 원내감염(healthcare-acquired infection, antimicrobial- resistant bacteria) 같은 감염질환의 감염예방 중재법을 평가하는 연구는 완수하기가 어렵고, 연구 예산도 부족하여 임상연구가 부족하다. 그러므로 감염예방정책의 근거는 매우 제한적일 수밖에 없고, 강력한 과학적 기반 없이 즉, 임상적/경제학적 영향을 어떻게 평가할지에 대한 분명한 틀 없이 상당한 자원이 임의로 투입되는 경우가 많다.

하지만, 효과적으로 환자를 보호하기 위해서는 다른 감염예방 중재법과의 철저한 비교효과평가가 수행되어야 하며, 이를 위해 군집 무작위대조 임상시험, 준 실험적 연구, 수학적 모델의 방법을 활용할 수 있다.

6.2.2. 방법

6.2.2.1. 군집 무작위대조 임상시험

군집 무작위대조 임상시험은 병원 전체, 즉 병원단위 무작위 배정 임상시험을 말하며, 복잡하고 비싸고 시간이 많이 소요된다. 하지만, 인구단위 중재법이 적절히 평가되기 위해서는 절대적으로 필요하다. 예를 들어, MRSA 검사 프로그램은 MRSA 보균자와 MRSA 전파를 예방하기 위해 격리된 환자를 모두 검사해야하며, 이런 검사 프로그램은 격리되지 않은 환자에게 간접적 이득이 있다. 이러한 인구단위 중재의 비교효과평가에 군집 무작위대조 임상시험이 적합하다. 하지만, 군집 무작위대조 임상연구는 질병의 발생 등 환자 안전의 관점에서 봤을 때 실행불가능하고 비윤리적일 수도 있어 외부 의뢰자의 통제나 연구변경 요청이 있을 수도 있으며, 종종 실현 불가능할 수도 있다.

4) Perencevich EN, Lautenbach E. Infection Prevention and Comparative Effectiveness Research. JAMA. 2011 Apr 13;305(14):1482-3.

6.2.2.2. 준 실험적 연구

준 실험적 연구는 무작위대조 임상심험과 군집 무작위대조 임상시험의 대안적 연구설계이며, 실험군이 스스로 대조군이 되어 중재 전후의 관심 있는 임상성과의 변화를 관찰하게 된다. 효율성, 일반화, 비무작위 임상성과 간의 일시적 연관성 평가의 장점이 있지만, 통제되지 않는 교란요인과 선택비뚤림 등 내적 타당도에 문제가 발생할 수 있어 통계분석의 주의를 요한다. 구간회귀분석(segmented regression analysis)은 중재 전후 감염율 추세의 변화를 평가하기에 적절한 방법인데, 독립변수를 몇 개의 구간으로 나누어 분리된 각 구간의 회귀선이 각 구간마다에 적합한 경우에 적용할 수 있다(김수경, 2010, Wagner 2002). 또한, 교란변수와 선택비뚤림을 통제하기 위해서 다변량분석(multivariable regression), 성향점수(propensity scores), 도구변수방법(instrumental variable methods)의 사용도 검토할 수 있다.

이외에도, 임상성과의 장기 평가나 중재를 하지 않는 비동등 대조군의 사용에서 한계가 있으며, 군집 무작위대조 임상시험과 준 실험적 연구 모두 유닛(unit)에 대한 중재가 평가되는데, 여기에 포함된 모든 개인의 동의를 받는 것이 불가능하다는 한계가 있다.

6.2.2.3. 수학적 시뮬레이션 모델

수학적 시뮬레이션 모델링 연구는 가용한 자료와 가정의 한계가 있지만, 가장 비용효과적인 연구방법이며, 예를 들어, 손씻기와 감염전파의 연관성을 통계적으로 추정하는 등에 활용될 수 있다.

6.3. 의료기기에 대한 비교효과연구[5)]

6.3.1. 개요

어느 것이 최적의 치료인지를 다른 기기와 비교하는 의료기기에 대한 임상시험은 눈가림(blinding)과 배정은폐(allocation concealment)가 어렵다. 한편, 관찰연구(observational study)의 경우에는 치료의 선택 이유 뿐 아니라 변경, 취소 사유 등을 알 수 없으며, 외과수술은 그 자체의 위험부담을 동반하기 때문에 수술 후 합병증과 실제 효과를 비교하기 위해서는 긴 추적관찰이 필요하다. 또한, 기기에 대한 안전성과 효과성을 평가할 때는 기기 자체의 특성 뿐 아니라 인체에 적용되었을 때 나타나는 영향(배터리 지속시간, 이식물의 물리적 마모 등)까지 평가되어야 하며, 대부분의 이식형 의료기기는 시간이 지남에 따라 계속 개선되므로 평가 시 이에 대한 고려도 필요하다.

6.3.2. 방법

6.3.2.1. 시술자 및 시술

의료기기 임상시험의 성과는 시술자의 능력과 숙련기간(operator's skill and learning curve)과 밀접하게 관련 있다. 대부분의 임상연구는 이미 숙련된 시술자에 의해 수행되는 것이 일반적이지만, 시술 방법에 따라 숙련에 필요한 시간이 다르기 때문에 이를 평가하는 것도 어렵고, 시술자의 선호도에 영향을 미치는 요인은 훈련, 병원이나 전문가집단의 표준, 급여, 처방에 대한 판단 등이 있다. 또한, 심근관통 혈관재형성(transmyocardial revascularization)의 예처럼 시술이나 중재 자체 뿐 아니라 그와 관련된 술기 및 접근 방법, 이미지의 양상(imaging modalities), 해부학적 부위(anatomic area) 등이 관련된 임상성과의 해석에서 고려해야 하며, 질병의 위중 정도에 따른 시술의 선택, 치료방법의 침습도 등도 감안되어야 하며, 중재를 시행하는 병원의 접근성과 함께 특정 시술을 제공하는 병원의 의도(예. 상업적

5) Sedrakyan A, Marinac-Dabic D, Normand SL, Mushlin A, Gross T. A Framework for Evidence Evaluation and Methodological Issues in Implantable Device Studies. Med Care. 2010 Jun;48(6 Suppl):S121-8.

또는 구조적 원인으로 특정 시술을 선호)도 중요하다.

6.3.2.2. 환자

임상시험 단계에서는 매우 제한된 환자들에게 시술이 되지만, 실제 임상현장(real-world setting)에서는 보다 많은 다양한 환자에게 시술이 되고("off-label" use), 이 결과들이 임상연구의 결과에 영향을 미친다. 또한, 대상 환자의 동반질환과 질환 중증도에 따라 영향을 받으며, 시술자의의 숙련도가 좋아질수록 질환 중증도에 의한 제한이 줄어들게 된다. 그러나 환자 선택요인(patient selection factors)은 임상자료나 등록자료로는 파악하기 어려우며, 특히 삶의 질과 같은 임상성과 등의 자료를 수집하기 어렵다.

6.3.2.3. 임상성과

임상적으로 의미 있는 임상성과를 측정하길 원하지만, 경우에 따라 대리결과변수(surrogate end-points)만 측정 가능하여 이로써 결론을 내기 어려운 경우도 있고, 소규모 연구로 인해 "combined critical end point", "composite end point"를 사용하게 되어 잘못된 결과를 유추할 수도 있다.

6.3.2.4. 문제점 및 해결방안

의료기기 임상시험의 눈가림, 의사의 숙련기간 차이 등의 문제점과 이에 대한 해결방안은 아래 표와 같다.

표 6-4 의료기기 임상시험의 문제점 및 해결방안 요약

문제점	추천하는 해결방법
눈가림 : 이중 또는 단일 눈가림이 어렵거나 불가능하다 (대부분의 RCT*, OS**)	강력한 배정은폐 방법(RTC) 독립적 임상성과 평가(RCT, OS)
무작위 배정의 시간 : 선정/제외가 무작위 배정 후 시술 개시에 영향을 받는다.(RCT, OS)	가능한 시술 교차에 대한 주의깊은 평가(RCT, OS) 모든 연구에서 intention-to-treat의 적용성과 한계를 논의(RCT, OS)
숙련기간 : 이식형 의료기기의 사용의 숙련도가 임상성과에 영향을 미칠 수 있다(RCT, OS) 숙련기간 : 이식형 의료기기의 사용의 숙련도가 임상성과에 영향을 미칠 수 있다(RCT, OS)	숙련된 전문가를 활용하여 비교 근거를 추정(RCT) 실제 임상현장에서 기술이 다양한 시술자를 등록시키는 것이 필요함(OS)
	가능하다면 숙련도, 수술건수와 진료결과의 상관관계(Volume-Outcome Relationship)과 수술건수의 역치를 보고(RCT, OS)
소수의 샘플 : 대부분 작은 차이를 발견하기에는 부족한 검정력과 일반적이지 않은 임상성과의 평가가 현실이다(RCT, OS)	부족한 검정력에서도 약간의 차이라도 발견할 수 있도록 모든 임상성과를 주의깊게 기록하고 보고(RCT, OS) 추적관찰기간이 긴 연구를 수행하고 베이지안 방법의 사용을 고려함(RCT, OS) 한 개 이상의 연구가 있다면 체계적 문헌고찰과 메타분석을 수행함(RCT, OS)
치료그룹에서의 공변량의 차이 : 환자/의사/병원의 차이가 크고 무작위배정이 어렵다(RCT, OS)	환자특성의 불균형을 모두 기록(RCT, OS) 알려진 교란변수를 보정할 수 있는 강력한 통계방법의 사용(RCT, OS) 방법의 선택, 추정의 타당성을 논의하고 잇점과 제한점을 목록화(RCT, OS)
기기 개선 : 연구진행중에서도 개량된 기기가 나올 수 있다(RCT, OS)	기기의 특성과 연구도중 변화가 있더라도 연구에서 여전히 유효한 임상성과를 정의함(RCT, OS)
일반화 : 아픈 환자의 제외하고 엘리트 시술자와 시술건수가 많은 병원을 포함함에 따라 임상시험에서 영향을 준다(RCT, OS)	연속된 환자와 다양한 경험과 훈련을 가진 시술자를 포함한 "All comer" 연구를 해야함(RCT, OS) 모든 이식형 의료기기와 관련된 변수를 표준 정의하려는 노력이 필요함(OS) 기기 식별자 필요(OS)

*RCT : 무작위배정 임상시험, **OS : 관찰연구

Reproduced from A Framework for Evidence Evaluation and Methodological Issues in Implantable Device Studies, 48(6 Suppl), Sedrakyan et al, Medical Care, S121-128, Copyright (2010), with permission from Lippincott Williams & Wilkins.

6.4. 안전성에 대한 비교효과연구

6.4.1. 개요

환자 안전성에 관한 연구와 비교효과연구는 공통적으로 약품(pharmaceuticals)의 효과와 이용이라는 주제를 다룬다. 과거 시장에 출시된 약품에 대한 연구는 주로 약품 출시 당시에는 발견하지 못했던 희귀한 부작용을 확인하는 데 초점이 맞추어져 있었다. 그러나 최근 연구에 따르면 약물에 의한 질병은 대개 발견되지 않은 희귀한 부작용이 아닌 일반적인 복용량과 연관된 부작용에 의한 것이다. 또한, 약물 출시 시점에 제공되는 정보들은 대개 약품의 효능(efficacy)에 대한 것이고, 약품의 효과(effectiveness), 특히 기존의 약물과 비교한 비교효과에 대한 정보는 충분하지 않다(Strom, 2007). 미국에서는 외래환자의 약물 처방에 대한 Medicare Part D 프로그램이 출범하면서, 약물의 효과와 안전성을 효과적으로 평가할 수 있는 방법에 대한 관심이 높아지고 있다(Lohr, 2007).

6.4.2. 방법

6.4.2.1. 연구방법

약물의 안전성과 효과를 평가하기 위한 다양한 방법들이 제시되고 있다. 우선 연구 설계 측면에서 임상시험이나 관찰연구 외에 군집 무작위배정 임상시험(cluster randomized trials)이나 실용시험(pragmatic trials) 또는 지연된 연구방법(delayed design)과 같은 다양한 접근법이 제시되었다. 군집 임상시험은 무작위배정이 의사나 치료법, 의료기관의 종류, 또는 지역으로 정의될 수 있는 집단을 기준으로 행해진다. 이 방법은 약물의 효과를 연구하는데 있어서 많은 이점을 제공한다. 실용시험의 경우 일반 환자나 치료법에 대해 적용이 가능하며, 보험과 수당에 대한 정책결정에 유용한 정보를 제공한다. 지연된 방법은 지역이나 의사들의 치료법을 군집으로 하는 무작위 임상 연구에서 제안되었다. 이는 일부 연구 대상 집단의 경우 일정 시간이 지난 후에만 중재법을 시행할 수 있도록 한 접근법으로, 지연된 시간은 유용한 도구변수(instrumental variable)로 사용될 수 있다.

약물의 효과와 환자의 안전성 연구는 다양한 자료원의 자료를 통해 약물 사용의

지속성, 순응도(adherence), 약물에 대한 결과 및 부작용과 같은 정보를 분석하여 수행된다. 사용 가능한 자료원으로는 Medicaid나 병원기록 등이 있다. 지금까지 자료의 타당성 평가나 시간의 흐름에 따른 약물 사용의 변화를 예측하기 위한 모델을 개발하기 위해 다양한 연구들이 시행되었다. 또한, 시뮬레이션 기술과 같은 새로운 방법들이 약물의 효과와 경제성, 그리고 안전성을 평가하기 위해 도입되었다(Lohr, 2007).

마지막으로, 약물의 부작용을 모니터하기 위해 지속적 혹은 주기적으로 개량된 신호를 감지하는 특성화된 감시체계(surveillance system)가 도입될 수 있다. 이러한 감시 체계는 약물 부작용을 조기에 감지할 수 있게 해주며, 시간에 따른 변이를 파악하는데 도움을 준다.

6.4.2.2. 고려 사항

환자의 안전성과 비교효과 연구에서 나타날 수 있는 방법론적인 문제점으로는 선택 바이어스, 노출 변수나 결과 변수의 오분류(misclassification)와 같은 계통적 오차(systematic error)를 들 수 있다. 이 외에도 랜덤 오차(random error)나 다른 약물 혹은 질병이나 약물 지표(drug indication)에 의한 교란효과 역시 문제점이 될 수 있다. 마지막으로, 데이터 접근 문제나, 특정 중재법(intervention)에 대한 연구의 어려움, 임상시험심사위원회(institutional review board, IRB)의 승인을 얻는데 있어서의 어려움과 같은 연구 수행 시 발생할 수 있는 문제점들이 있다(Strom, 2007).

노출인자의 오분류(exposure misclassification)는 장기적인 약물의 효과 연구 시 고려되어야 할 사항이다. 과거 약물 사용의 패턴이나 사용한 약물의 종류는 환자의 기억에 의존하게 되는데 환자들은 종종 그들이 사용하는 약물을 정확히 알지 못하며, 약물 사용에 대한 환자의 기억의 유효성을 판단할 수 있는 정보는 거의 존재하지 않는다. 호르몬 연구로부터 얻어진 데이터들에 따르면 환자의 회상(recall)은 약물의 종류나, 사용한 기간, 얼마나 최근에 사용했는지, 그리고 환자의 인구학적 특성에 따라 달라진다.

복용된 약물의 양에 대한 오분류는 약물 사용을 추적한 보험청구 자료를 이용할 경우에 발생할 수 있다. 특히 액체나 기체 타입의 약물일 경우나, 필요할 시 복용과 같은 처방의 경우 오분류는 더 자주 발생할 수 있다. 또한, 환자의 순응도나 지속성과 같은 문제가 제기될 수있다.

역학이나 보건서비스 연구는 주로 이분적인 노출 변수에 대해 이루어졌지만, 약

물의 경우는 이분화가 되는 경우가 드물다. 또한, 중요한 약물의 부작용은 주로 복용량과 관련된다. 최고치와 최소치 복용량의 범위가 큰 약물의 경우는 효과의 범위 역시 크다. 부작용은 사용 기간과도 관련이 있다. 과민증과 같이 부작용이 급성으로 발생하는 경우가 있는가 하면, 암발생과 같이 장기간 축적된 복용에 의해서 나타나기도 한다. 또한 새로운 사용자와 기존의 사용자 사이에서도 차이가 나타날 수 있으며, 마지막 복용과 관련된 부작용의 패턴 역시 다양하다. 예를 들어, 부작용이 처음 복용시에만 발생할 수도 있고, 매 복용시 발생할 수 있으며, 복용을 중단하면 부작용이 사라지는 경우도 있다.

결과변수에 대한 오분류(outcome misclassification) 또한 연구의 타당성을 해칠 수 있다. 많은 연구들이 임상적 결과를 연구하는 대신 중간(intermediate) 결과를 연구한다(예. 심근경색의 경우는 혈압). 특히, 약물이 시판되기 전 시행된 연구들은 대리지표(surrogate outcomes)에 대해 연구하는 경향이 있으며, 이는 결과 오분류를 일으킨다. 진통제와 같은 경우 결과를 측정하기 어렵고, 약한 발진의 경우는 청구자료에 포함되지 않기 때문에 정확한 결과를 예측하기가 어렵다. 또한, 진단 결과의 타당성이 불확실한 경우 문제가 되는데, 청구자료를 사용할 경우 특히 문제가 될 수 있다. 입원환자의 병원청구 진단 자료는 퇴원시 진단과 잘 상응한다. 반면, 외래환자의 임상진단은 타당성을 입증하기 어렵다. 따라서, 연구자들은 진단에 대한 타당성을 확보하기 위해 의료 기록을 정기적으로 확보하는 것이 필요하다.

교란효과는 일반적으로 무작위배정이나 제한(restriction) 및 매칭과 같이 설계 방법에서 제어되거나, 층화나 다변수모형을 이용한 분석 단계에서의 통제가 이루어질 수 있다. 대부분의 방법이 약물 효과 연구에서 유용하지만, 무작위배정을 제외한 나머지 방법에서는 교란요인의 측정이 필요하다. 그러나 교란요인의 측정은 주요 교란요인이 다른 약물이나 질병인 경우가 대다수이므로, 노출인자나 결과 변수 측정과 마찬가지로 여러 어려움이 존재한다. 특히 약물의 효과에 대한 연구에서 적응도에 의한 교란은 중요한 문제점이다. 약을 처방받은 사람과 그렇지 않은 사람 사이에는 차이점이 존재하는데, 이러한 지표는 대개 결과와도 연관이 있다. 이 지표가 확실하게 측정되지 않으면 통제되지 않은 교란효과와 선택 바이어스가 발생할 수 있다. 예를 들어, 환자의 치료전 혈압을 측정하는 것은 쉽지만 고혈압 약을 선택하는 것은 추가적인 요소(예. 부작용에 대한 환자의 저항력(tolerability))의 영향을 받는다. 또한, 시간당 복용량과 약의 복용을 지속할지 여부에 대한 선택은 환자의 혈압반응치 및 다른 부작용의 유무에 의해 결정된다. 이러한 상황에서 시간에 따른 변화를 반영한 적응도에 의한 교란을 측정하는 것은 매우 어렵다.

환자의 안전과 비교효과 연구에서는 데이터 접근과 같이 실행과 관련된 문제점

이 제기될 수 있다. 예를 들어, 다른 Medicare 청구자료와 연계되었을때 Medicare part D는 중요한 자료원이 될 수 있지만, 연구집단에 대한 접근이 보장되지 않는다는 단점이 있다. 환자 안전성 연구에 있어서도 지역보건관리 시스템에 접근하면 연구진행이 용이하나, 이러한 접근은 지역 통제나 데이터의 질적 수준, 동료검증(peer review) 보호 등과 같은 문제점이 야기될 수 있다. 또한, 대부분의 청구자료들은 진단의 타당성을 위해 진료자료를 필요로 하지만, 진료기록을 얻기 위해서는 비용 문제나 IRB의 승인과 같은 어려움이 있다. 연구 수행시 발생할 수 있는 또 다른 문제점으로는, 환자 안전성이나 비교효과 연구의 필요성을 사람들에게 인지시키는 것이다. 특히 정보기술(information technology)을 이용한 중재법의 경우 안전성과 효과연구의 체계화에 도움을 주지만, 일부 중재법은 해를 끼칠 수 있거나 전혀 이익을 주지 않기도 한다. 따라서 IT를 이용한 중재법을 평가하는데 어려움이 존재한다. 마지막으로 IRB 승인에 대한 문제점이 야기될 수 있다. 우선 지역적으로 분산되어 있는 병원의 기록이나 환자들을 이용한 연구의 경우 IRB의 승인을 받기가 어려울 수 있다. 또한, 임상시험이나 IT를 이용한 중재법에 대한 연구의 경우 비윤리적 문제가 발생할 우려가 있으며, 이로 인해 IRB 승인을 받지 못할 수 있다(Strom, 2007).

6.4.3. 사례

Confavreux 등(2001)은 다발성 경화증(multiple sclerosis, MS)환자를 대상으로, 환자-교차설계 연구를 이용하여 B형 간염 및 다른 백신과 MS 재발의 상관관계를 조사하였다. 최초 MS 재발 전 2개월이 위험기간(risk period)이며, 위험 기간 전 두 달씩 4번의 기간이 대조기간(control period)으로 설정되었다. 각각의 기간 동안 백신 접종이 이루어진 환자들은 노출집단으로, 그렇지 않은 환자들은 비노출집단으로 분류되었다. 분석 결과, B형 간염 및 다른 백신 접종과 단기간 다발성 경화증의 재발 사이에는 유효한 상관관계가 나타나지 않았다(Confavreux 등, 2001).

참고문헌

김수경, 김희은, 백미숙, 이숙. 급성상기도감염 항생제 처방률 공개 효과 분석. Kor. J. Clin. Pharm., Vol. 20, No. 3. 2010.

현민경, 장은진, 박주연, 안지혜. 류마티스 관절염 환자에서 항 CCP 항체 검사의 임상적 유용성 평가. 한국보건의료연구원 2012

Chang SM, Matchar DB, Smetana GW, Umscheid CA, editors. Methods Guide for Medical Test Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Jun.

Confavreux C, Suissa S, Saddier P, et al. Vaccinations and the Risk of Relapse in Multiple Sclerosis. New England Journal of Medicine. 2001;344:319–326.

Gluud C, Gluud LL. Evidence based diagnostics. BMJ. 2005;330(7493):724–6.

Lohr KN. Emerging Methods in Comparative Effectiveness and Safety: Symposium Overview and Summary. Medical Care. 2007:45;S5–S8.

Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.

Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data‐.analytic approaches and some additional considerations. Stat Med 1993; 12: 1293–1316.

Perencevich EN, Lautenbach E. Infection Prevention and Comparative Effectiveness Research. JAMA. Apr 2011;305(14):1482–3.

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982–990.

Rutter CM, Gatsonis CA. Regression methods for meta‐.analysis of diagnostic test data. Acad Radiol 1995; 2 Suppl1:S48–56.

Sedrakyan A, Marinac–Dabic D, Normand SL, Mushlin A, Gross T. A Framework for Evidence Evaluation and Methodological Issues in Implantable Device Studies. Med Care. 2010 Jun;48(6 Suppl):S121–8.

Strom BL. Methodologic Challenges to Studying Patient Safety and Comparative Effectiveness. Medical Care. 2007:45;S13–S15.

Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther. 2002 Aug;27(4):299-309

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, et. al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36

용 | 어 | 정 | 리

2상/3상 통합 임상시험 설계	adaptive seamless II/III trial phase
가설 조정 설계	hypothesis-adaptive design
간접비교	indirect treatment comparison
관찰연구	observational study
건강수명	Healthy Years Equivalents, HYE
결정분석 모형	decision analytic mode
결정수형 모형	decision tree
경시적 연구	longitudinal study
경시적 자료	longitudinal data
계층모형	hierarchical model
계층적 자료	hierarchical data
계통적 오차	systematic error
곡선하 면적	Area Under the Curve, AUC
공통대조군	common comparator
구간회귀분석	segmented regression analysis
군집 무작위 대조 임상시험	cluster randomized trial
교란변수	confounding
교환가능성	exchangeability
그룹 축차 설계	group sequential design
근거 네트워크	network of evidence
깔대기그림	funnel plot
네트워크 메타 분석	network meta-analysis
다중 조정 설계	multiple adaptive design
대리지표	surrogate marker
동질성	homogeneity
마콥 모형	Markov model
메타분석	meta-analysis
무작위 배정 임상시험	Randomized Controlled Trials, RCT
무작위 배정 조정 설계	adaptive randomization design
민감도분석	sensitivity analysis
배정된 대로 분석하는 방법	intention-to-treat, ITT
베이지안 방법	Bayesian approach
변경률	switchback rate
비모수 조정 설계 방법	non-parametric adaptive urn design approach
비열등성 가설	non-inferiority hypothesis

비열등성 허용 한계	non-inferiority margin
비용-최소화 분석	cost minimization analysis
비용-편익 분석	cost-benefit analysis
비용-효과 수용 곡선	cost-effectiveness acceptability curve
비용-효과분석	cost-effectiveness analysis
비용-효용 분석	cost-utility analysis
사전분포	prior distribution
사전정보	prior information
사후분포	posterior distribution
선형혼합모형	linear mixed model
성향 점수	propensity score
순건강편익	net health benefit
숲그림	forest plot
실용임상시험	Pragmatic Clinical Trial, PCT
실험별 제1종 오류	experiment-wise type I error
약동학	pharmacokinetics
약물 용량 조정 설계	adaptive dose findings design
양성예측도	positive predictive value
양성우도비	likelihood ratio of positive predictive value
역확률 가중치	Inverse Probability of Treatment Weighting, IPTW
연속적 재평가 방법	Continual Re-assessment Method, CRM
열등 반응 탈락 설계	drop-the-loser design
용량 확정 설계	dose finding design
우도비검정	likelihood ratio test
우위성 가설	superiority hypothesis
위약-대조 시험	placebo-controlled trial
유사성	similarity
음석예측도	negative predictive value
음성우도비	likelihood ratio of negative predictive value
이변량 랜덤효과모형	bivariate random effect model
이질성	heterogeneity
일관성	consistency
일반화 랜덤효과 모형	generalized random effects model
일반화 선형모형	generalized linear model
일반화 선형혼합모형	generalized linear mixed model
일반화 추정방정식	generalized estimating equation
자료검토위원회	Data monitoring Committee, DMC
자연 단위	natural unit

작업상관행렬	working correlation matrix
잔차 교란	residual confounding
장애모수	nuisance parameter
적응임상시험	adaptive clinical trial
적응증에 의한 교란	confounding by indication
전이모형	transitional model
전자 건강 기록	electronic health record
점증적 비용-효과비	Incremental Cost-Effectiveness Ratio, ICER
정량적 합성	quantitative synthesis
정성적 합성	qualitative synthesis
조건부모형	conditional model
주변모형	marginal model
중간분석	interim analysis
직접비교	head to head
직접비교 효과시험	head-to-head effectiveness trial
진단검사	diagnostic test
진단법 문헌고찰	diagnostic literature review
진단오즈비	diagnostic odds ratio
질보정수명	Quality Adjusted Life Years,QALY
최대내성용량	Maximum Tolerable Dose, MTD
최소유효량	Minimum Effective Dose, MED
치료 전환 설계	adaptive treatment-switching design
특이도	specificity
평균 비용-효과비	Average Cost-Effectiveness Ratio, ACER
표본수 재추정 설계	sample size re-estimation design
학습단계	learning stage
혼합비교	mixed(multiple) treatment comparison
확인단계	confirmatory stage
환자-교차설계 연구	case-crossover study
환자-시간-대조군 연구	case-time-control study
활성 대조군	active control group
횡단면연구	cross-sectional study
효과변경인자	effect modifier
효과변경작용	effect modification
ROC커브	Receiver-Operating Characteristic curve
SROC커브	Summary Receiver-Operating Characteristic curve

색 | 인

비매품
93510
9 788968 340024
ISBN 978-89-6834-002-4
ISBN 978-89-6834-003-1 (세트)

NECA 한국보건의료연구원
National Evidence-based Healthcare Collaborating Agency