



### 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)



위내시경 영상 분석 인공지능  
소프트웨어의 안전성 및 유효성 평가를  
위한 임상시험 프로토콜 개발

연세대학교 대학원  
의료기기산업학과  
최희영

위내시경 영상 분석 인공지능  
소프트웨어의 안전성 및 유효성 평가를  
위한 임상시험 프로토콜 개발

지도교수 구 성 욱, 장 원 석

이 논문을 석사 학위논문으로 제출함

2022년 12월

연세대학교 대학원

의료기기산업학과

최 희 영



# 최희영의 석사 학위논문을 인준함

심사위원 구성욱 인

심사위원 장원석 인

심사위원 김지현 인

연세대학교 대학원

2022년 12월

## 감사의 글

어느덧 석사 졸업을 앞두고 있다니 감회가 새롭습니다. 학부생부터 지금까지 2년이 넘는 시간 동안 학문적으로 많이 배웠을 뿐만 아니라 인간적으로도 성숙할 수 있던 시간이었습니다. 그 마지막 결과물로 졸업 논문을 완성하며, 도와주신 모든 분들께 진심으로 감사의 인사를 남기고 싶습니다.

아낌없는 조언과 가르침 주신 장원석 교수님, 바쁘신 와중에도 연구와 논문에 대해 세심하게 지도해 주시고 이끌어주셔서 감사합니다. 학문적 발전을 위해 성심성의껏 지도해 주신 구성욱 교수님, 함께 프로토콜 개발에 힘써주시고 임상적으로 도움을 주신 김지현 교수님께도 깊은 감사를 드립니다.

긴 시간 동안 연구실에서 함께 연구하고 수학하며 힘이 되어준 연구실 선생님들, 아낌없는 응원과 격려를 보내준 친구들과 지인들에게도 감사의 말을 전합니다. 마지막으로, 믿고 지켜봐 주시고 늘 곁에서 진심 어린 관심과 애정을 아끼지 않으셨던 부모님께 무한한 감사를 드립니다.

많은 사람들 덕분에 잊지 못할 따뜻한 대학원 생활을 했습니다. 고마운 마음 잊지 않고 보답하겠습니다. 다시 한번 감사의 마음을 전하며, 졸업이 끝이 아닌 또 다른 출발점이라 생각하고 더욱더 발전하고 성장하여 의료기기산업에 제 역량을 베풀 수 있도록 노력하겠습니다.

최희영 올림

## 차 례

그림 차례 .....	iii
표 차례 .....	iv
국문요약 .....	vii
<b>I. 서론 .....</b>	<b>1</b>
1. 연구 배경 .....	1
2. 연구 목적 .....	5
<b>II. 재료 및 방법 .....</b>	<b>7</b>
1. 위암의 임상적 이해 .....	7
2. 영상검출·진단보조 소프트웨어의 이해 .....	13
가. 국내 영상검출·진단보조 소프트웨어 허가 현황 .....	15
나. 의료기기 소프트웨어 관련 규제 .....	20
다. 영상검출·진단보조 소프트웨어 임상시험의 이해 .....	24
3. 국내외 위암 인공지능 소프트웨어 성능 평가 및 임상시험 사례 ..	28
가. 해외 소프트웨어 성능 평가 사례 .....	28
나. 인공지능 소프트웨어 임상시험 설계 사례 .....	29
4. 임상시험 대상 의료기기 개요 .....	49
<b>III. 결과 .....</b>	<b>52</b>
1. 위내시경 영상 분석 인공지능 소프트웨어 임상시험 방법 .....	52
가. 임상시험 설계 .....	52



나. 표본 데이터 선정 .....	58
2. 위내시경 영상 분석 인공지능 소프트웨어 임상적 유효성 평가 .....	65
가. 유효성 평가 분석군 .....	65
나. 일차 유효성 평가변수 .....	66
다. 이차 유효성 평가변수 .....	68
3. 위내시경 영상 분석 인공지능 소프트웨어 임상적 안전성 평가 .....	72
<b>IV. 고찰 .....</b>	<b>73</b>
<b>V. 결론 .....</b>	<b>75</b>
<b>참고문헌 .....</b>	<b>76</b>
<b>Abstract .....</b>	<b>79</b>

## 그림 차례

그림 1. 연도별 신규 발생 암 환자 수	3
그림 2. 2019년 암종별 발생 빈도	3
그림 3. 2019년 남성 암 발생 빈도	4
그림 4. 2019년 여성 암 발생 빈도	4
그림 5. 2020년 전 세계 암 신규 발생 수	7
그림 6. 위의 해부학적 분류	8
그림 7. 진행성 위암의 육안적 분류	10
그림 8. T병기 도식화	12
그림 9. ROC curve 및 AUC 그래프	27
그림 10. 위내시경 영상 분석 화면	49
그림 11. 소프트웨어 구성	50
그림 12. PC와 소프트웨어 연동 화면	51
그림 13. 우월성 검정 신뢰구간	54
그림 14. 비열등성 검정 신뢰구간	54
그림 15. 동등성 검정 신뢰구간	55

## 표 차례

표 1. 위암영상검출·진단보조소프트웨어의 정의 .....	6
표 2. 육안적 분류 .....	9
표 3. T병기 .....	11
표 4. 영상검출·진단보조 소프트웨어 품목분류 .....	13
표 5. 국내 영상검출·진단보조 소프트웨어 허가 현황 .....	15
표 6. 소프트웨어 관련 국외 규격 .....	21
표 7. 소프트웨어 관련 국내 가이드라인 .....	22
표 8. 성능 평가 항목 정의 및 계산식 .....	25
표 9. 이분형 검사 결과 요약 .....	26
표 10. 해외 인공지능 소프트웨어 성능 평가 리스트 .....	28
표 11. 해외 인공지능 소프트웨어 임상시험 리스트 .....	29
표 12. 해외 인공지능 소프트웨어 임상시험 설계 사례 1 .....	30
표 13. 해외 인공지능 소프트웨어 임상시험 설계 사례 2 .....	31
표 14. 해외 인공지능 소프트웨어 임상시험 설계 사례 3 .....	32
표 15. 해외 인공지능 소프트웨어 임상시험 설계 사례 4 .....	33
표 16. 해외 인공지능 소프트웨어 임상시험 설계 사례 5 .....	34
표 17. 국내 인공지능 소프트웨어 임상시험 리스트 .....	35

표 18. 국내 인공지능 소프트웨어 임상시험 설계 사례 1	37
표 19. 국내 인공지능 소프트웨어 임상시험 설계 사례 2	38
표 20. 국내 인공지능 소프트웨어 임상시험 설계 사례 3	39
표 21. 국내 인공지능 소프트웨어 임상시험 설계 사례 4	40
표 22. 국내 인공지능 소프트웨어 임상시험 설계 사례 5	42
표 23. 국내 인공지능 소프트웨어 임상시험 설계 사례 6	44
표 24. 국내 인공지능 소프트웨어 임상시험 설계 사례 7	45
표 25. 국내 인공지능 소프트웨어 임상시험 설계 사례 8	46
표 26. 국내 인공지능 소프트웨어 임상시험 설계 사례 9	47
표 27. 평행설계 및 교차설계	53
표 28. 대조군 종류	55
표 29. 임상시험 디자인 변경 과정	57
표 30. 선정 및 제외 기준	59
표 31. 선행연구 민감도 및 특이도 결과	60
표 32. Clopper-Pearson의 정확신뢰구간	61
표 33. 내부 성능시험 결과	62
표 34. 민감도 우월성 검정 가설	62
표 35. 데이터 수 산출식	63
표 36. 특이도 우월성 검정 가설	63



표 37. 표본 테이터 수 .....	64
표 38. 위암 및 비위암 하위그룹 .....	65
표 39. IoU 계산식 .....	67
표 40. 정확도 계산식 .....	69
표 41. 병변의 위치별 해당 부위 .....	71
표 42. 이상사례 위해정도 .....	72

## 국문요약

### 위내시경 영상 분석 인공지능 소프트웨어의 안전성 및 유효성 평가를 위한 임상시험 프로토콜 개발

위암은 2019년에 갑상선암, 폐암의 뒤를 이어 세 번째로 가장 많이 발생한 빈도가 높은 질환이다. 위내시경은 위암을 확진할 수 있는 가장 정확한 검사로, 위 내부를 직접 관찰하면서 위암으로 의심되는 병변을 발견하고 조직진단을 시행할 수 있다. 소화기내시경 검사에서 내시경 의사들 간, 내시경 수련 과정에 따라 내시경 검사 성격에 유의한 차이가 있다는 연구들이 보고되면서 내시경 질 관리에 대한 관심이 고조되었다. 내시경 검사 수준의 편차를 줄이기 위해 내시경 영상 판독에 인공지능을 적용할 수 있다. 인공지능 소프트웨어가 위암 예측 확률(%)을 나타내고 병변 위치를 표시해 주어 내시경 검사 단계에서 의료진을 보조하고 의료진의 임상적 결정에 도움이 될 수 있다. 또한, 초기에 발견함으로써 위암 환자의 예후 개선에도 도움이 될 것으로 보인다.

본 연구에서는 위암과 위암의 검출 및 진단 소프트웨어 관련 문헌을 조사 및 분석하여 임상시험 프로토콜 개발 시 요구되는 규정 및 가이드라인을 바탕으로 위암영상검출·진단보조소프트웨어에 대한 임상시험 디자인과 유효성 평가 방법을 제시하고 있다. 제시한 프로토콜을 바탕으로 소프트웨어의 유효성과 안전성이 검증되어 위암을 진단함에 있어 의료진에게 도움을 주고, 향후 영상 분석 인공지능 소프트웨어 프로토콜 개발 시 활용되며, 인공지능에 기반한 영상검출 및 진단보조 소프트웨어가 활발히 연구개발되기를 기대한다.

---

핵심 되는 말: 위암, 위내시경, 임상시험, 인공지능, 소프트웨어, 민감도, 특이도

# 위내시경 영상 분석 인공지능 소프트웨어의 안전성 및 유효성 평가를 위한 임상시험 프로토콜 개발

<지도교수 구 성 옥, 장 원 석>

연세대학교 대학원 의료기기산업학과

## 최희영

### I. 서론

#### 1. 연구 배경

보건복지부와 중앙암등록본부에서 발표한 국가암등록통계 자료에 따르면 2015년 이후 신규 암 환자 수는 매년 증가하는 추세로 2019년에 약 25.5만 명이 발생하였다(그림 1).<sup>1</sup> 위암은 2019년에 갑상선암, 폐암의 뒤를 이어 세 번째로 가장 많이 발생하였으며(그림 2), 남성에서 발생률 2위(그림 3), 여성에서 발생률 4위(그림 4)를 차지한 빈도가 높은 질환이다.<sup>1</sup>

한국은 국내 사망원인 1위인 암 질환으로 인한 고통, 피해 및 사회적 부담을 줄이기 위해 국가암관리정책을 3차에 걸쳐 추진해왔고, 그 결과 암생존율이 1995년 42.9%에서 2018년 70.7%까지 향상되었다.<sup>1</sup> 2021년부터는 전 주기적 암관리 강화를 위한 제4차 암관리 종합계획을 수립하여 암 빅데이터 활성화, 암 예방·검진 고도화, 임 치료·대응 내실화 등의 목표를 수립하고 이행 중에 있다.<sup>2</sup> 또한, 국가암검진 프로그램을 추진하여 위암을 대상으로 위내시경 검사

를 시행함으로써 위암을 조기에 발견하여 치료율을 높이고 암으로 인한 사망을 줄이고자 지속적으로 모니터링하고 있다.

위암의 병리학적 병변 진단을 위해 면밀한 위장 검사, 병변 식별, 표적 생검이 중요하다. 위내시경은 위암을 확진할 수 있는 가장 정확한 검사로, 위 내부를 직접 관찰하면서 위암으로 의심되는 병변을 발견하고 조직진단을 시행할 수 있다.<sup>3</sup> 내시경은 침습적이고 성공적인 관찰이 쉽지 않은 전문적인 의료지식과 기술이 필요한 검사이고, 위 종양 검출은 내시경 전문의의 경험, 전문성, 숙련도에 따라 좌우되므로 정확하고 안전한 내시경 검사를 위해 내시경 전문의와 보조자의 역할이 매우 중요하다.<sup>3,4</sup> 소화기내시경 검사에서 내시경 의사들 간, 내시경 수련 과정에 따라 내시경 검사 성적에 유의한 차이가 있다는 연구들이 보고되면서 내시경 질 관리에 대한 관심이 고조되었다. 개인별, 기관별 내시경 검사 수준의 편차를 줄이기 위해 내시경 영상 판독에 있어 인공지능(Artificial Intelligence, AI)을 적용할 수 있다. 인공지능 소프트웨어가 위암 예측 확률(%)을 나타내고 병변의 위치를 표시해 주어 내시경 검사 단계에서 내시경 전문의를 보조하고 이들의 임상적 결정에 도움이 될 수 있다.<sup>5</sup> 내시경에 인공지능을 적용함으로써 위암의 내시경 치료 가이드라인을 제공하거나 침습 깊이를 예측하여 불필요한 수술을 피할 수 있을 뿐만 아니라 환자의 전반적인 예후 개선에도 도움이 될 것으로 기대된다.<sup>4</sup>

2018년 의료기기법이 개정되면서 의료기기 정의에 소프트웨어 항목이 포함되어 의료기기로 관리됨에 따라 위암영상검출·진단보조소프트웨어 품목이 신설되었다. 임상시험 설계 당시 해당 품목으로 허가받은 제품은 없었지만, 2022년 12월 기준 1건의 허가 제품을 확인하였다. 본 연구를 통해 적절한 유효성 평가변수 및 평가 방법을 제시하여 위암영상검출·진단보조소프트웨어의 안전성과 유효성을 증명하기 위한 임상시험 프로토콜을 제시하고자 한다.

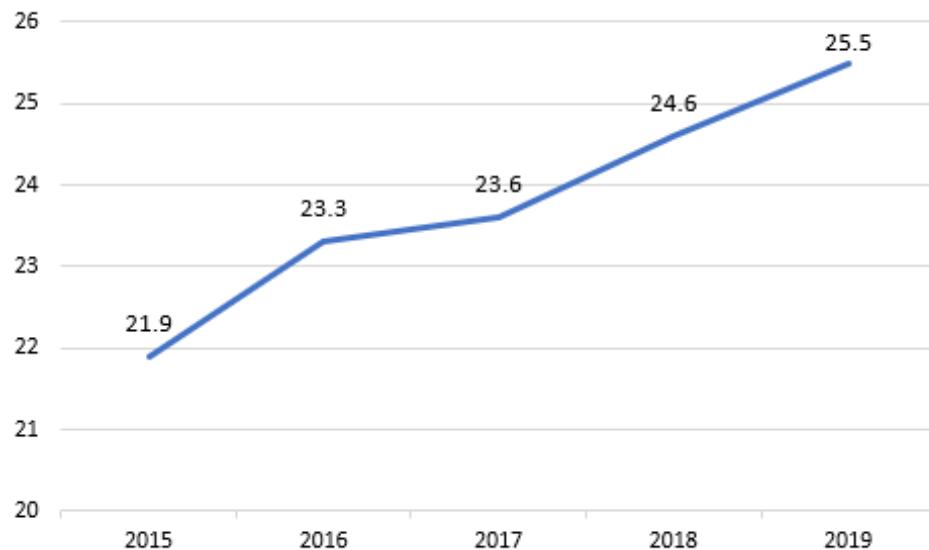


그림 1. 연도별 신규 발생 암 환자 수 (단위: 만 명)

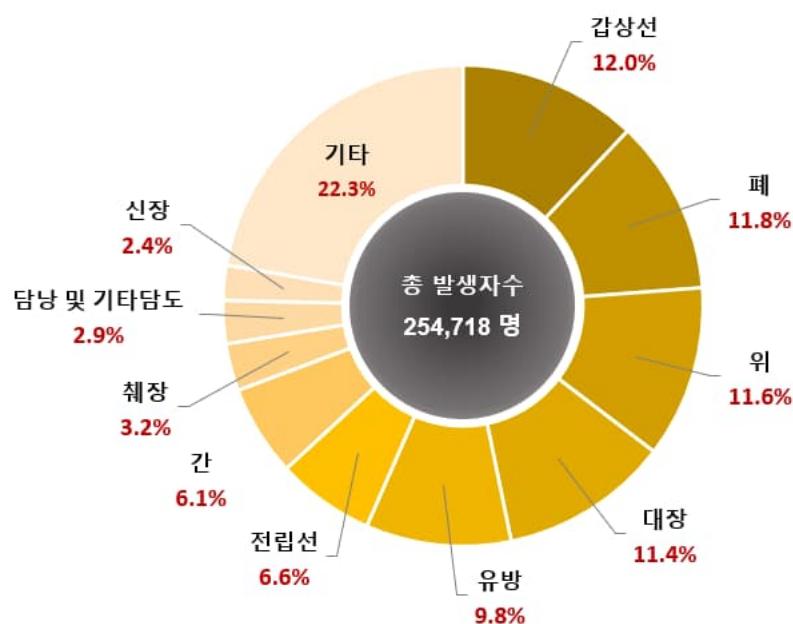


그림 2. 2019년 암종별 발생 빈도

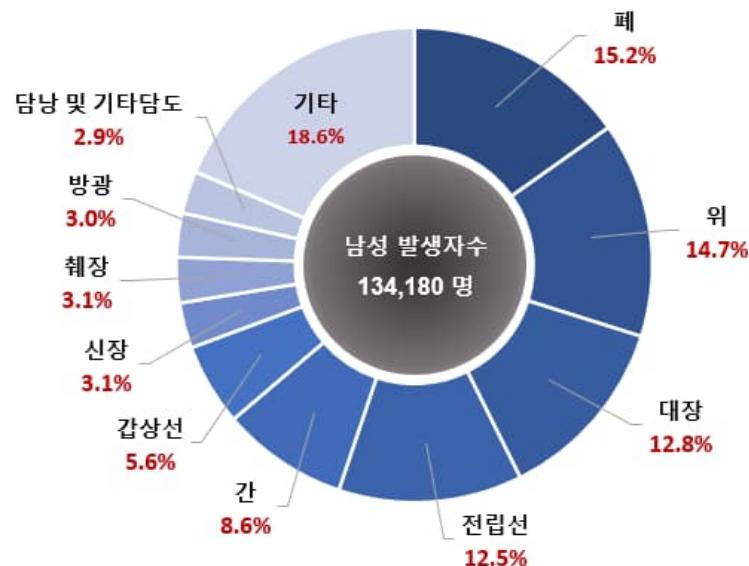


그림 3. 2019년 남성 암 발생 빈도

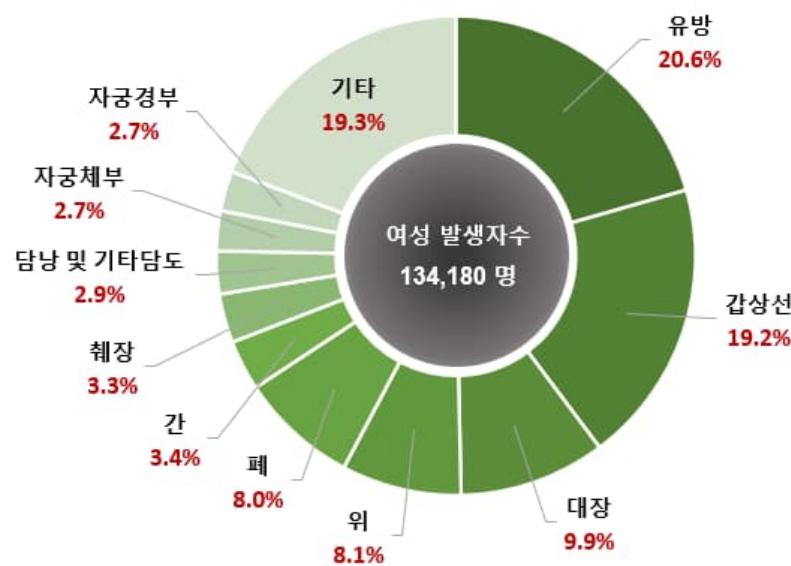


그림 4. 2019년 여성 암 발생 빈도

## 2. 연구 목적

위암영상검출·진단보조소프트웨어는 내시경영상 내에서 위암 의심부위를 검출한 후 윤곽선, 색상 또는 지시선 등으로 표시하거나 위암의 유무, 위암의 중증도 또는 위암의 상태 등에 대한 가능성 정도를 자동으로 표시하여 의료인의 진단결정을 보조하는 데 사용하는 소프트웨어이다(표 1). 본 연구에서 다루는 위암영상검출·진단보조소프트웨어로 허가받고자 하는 제품은 내시경 검사 장비로부터 획득한 위내시경 영상을 바탕으로 학습된 인공지능 모델을 사용한다. 딥러닝 기반의 콘볼루션신경망(Convolutional Neural Network, CNN)을 기반으로 개발되어, 학습데이터를 통해 영상의 특징을 학습하고, 학습된 내용을 바탕으로 내시경 영상을 분석하고 위암 의심 병변을 표시해 준다.

위내시경 검사에서 인공지능을 이용하여 자동으로 위암을 검출한 사례들을 조사하고, 선행연구에서 설계한 내용을 바탕으로 위암 유무를 진단하는 데 있어 위내시경 영상 분석 인공지능 소프트웨어의 안전성과 유효성을 평가하기 위한 임상시험 프로토콜을 개발하고자 한다.

위내시경은 위 내부를 직접 관찰하면서 증상이 없는 조기 위암(Early Gastric Cancer, EGC) 발견에 가장 좋은 검사이며, 위암이 진단된 경우 위암의 모양, 크기, 위치를 평가하고 수술 범위를 결정하는 데 필수적이다.<sup>3</sup> 위암의 국내 발생 빈도가 높고, 내시경 검사 질관리에 대한 관심이 고조되는 실정에서, 위내시경 검사를 통해 위암을 조기에 진단하여 치료율을 높이고 사망률은 낮추며, 위내시경 영상 분석 인공지능 소프트웨어가 의료진의 부담을 덜어주고 내시경 질관리를 높이는 데 기여하고자 한다.

본 연구에서는 위암과 위암의 검출 및 진단 소프트웨어 관련 문헌을 조사 및 분석하여 임상시험 프로토콜 개발 시 요구되는 규정 및 가이드라인을 바탕으로 위암영상검출·진단보조소프트웨어의 품목허가를 위한 임상시험 프로토콜



을 제시하며, 임상적 유효성과 안전성이 검증된 인공지능 소프트웨어를 통해 위암을 진단함에 있어 의료진에게 도움을 주고, 향후 영상 분석 인공지능 소프트웨어 프로토콜 개발 시 활용되며, 인공지능에 기반한 영상검출 및 진단보조 소프트웨어가 활발히 연구개발되는 데에 기여하고자 한다.

표 1. 위암영상검출·진단보조소프트웨어의 정의

품목명	위암영상검출·진단보조소프트웨어
영문 품목명	Gastric cancer image, computer aided detection/diagnosis software
품목코드	E04020.02
등급	3
정의	내시경영상 내에서 위암 의심부위를 검출한 후 윤곽선, 색상 또는 지시선 등으로 표시하거나 위암의 유무, 위암의 중증도 또는 위암의 상태 등에 대한 가능성 정도를 자동으로 표시하여 의료인의 진단결정을 보조하는데 사용하는 소프트웨어

출처: 의료기기 품목 및 품목별 등급

## II. 제료 및 방법

### 1. 위암의 임상적 이해

암은 인체 몸에 생기는 악성 종양이고, 악성 종양이 위에 생긴 것을 위암이라고 한다.<sup>6</sup> 2020년에 세계적으로 약 1,930만 명의 암 신규 환자가 발생하였다 (그림 5).<sup>7</sup> 유방암, 폐암, 대장암, 전립선암에 이어 5위를 차지한 위암은 약 109만 명의 신규 환자가 발생하였고, 약 77만 명이 위암으로 인해 사망하였다.<sup>7</sup> 이처럼 위암은 전 세계적으로 발병률이 높고, 암의 주요 사망 원인 중 하나이기 때문에 이를 예방하고, 조기에 검출하는 것이 중요하다. 위암은 일본내시경학회에서 발표한 육안적 분류에 따라 분류할 수 있으며, TNM 분류를 기반으로 종양의 침습 깊이(Depth of tumor invasion), 림프절 전이(Lymph node metastasis), 원격 전이(Distant metastasis)에 따라 위암의 병기를 분류할 수 있다.<sup>8</sup>

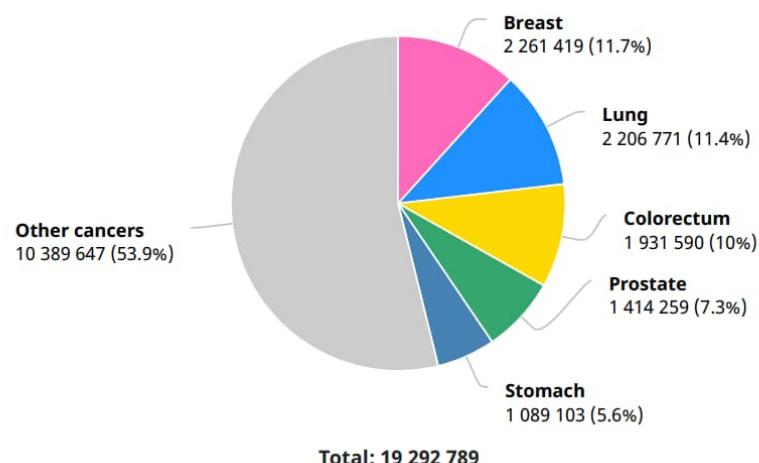


그림 5. 2020년 전 세계 암 신규 발생 수

위는 해부학적으로 소만과 대만의 삼등분 점을 연결하는 선에 따라 상부(Upper, U), 중부(Middle, M), 하부(Lower, L) 세 부분으로 구분된다(그림 6).<sup>8</sup> 상부는 식도와 연결되는 부위인 분문부, 위의 윗부분인 위저부, 위저부 아래 상부체부를 포함하고, 중부는 위저부 아래 중부체부와 하부체부, 위각부를 포함하며, 하부는 위체부 아래 전정부와 십이지장과 연결되는 부위인 유문부를 포함한다.<sup>9</sup>

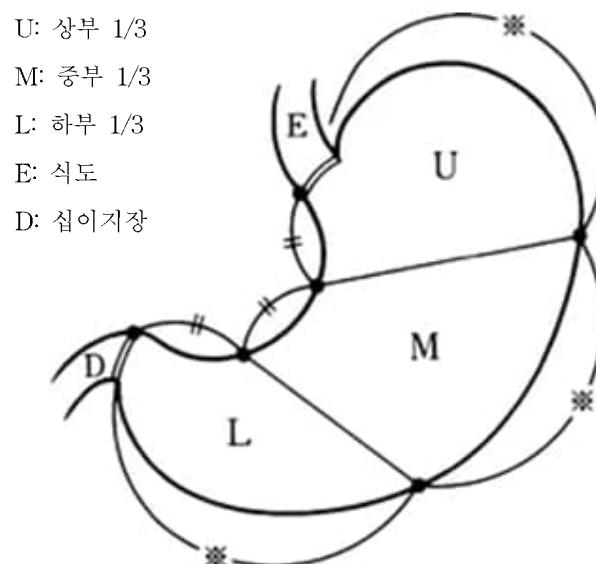


그림 6. 위의 해부학적 분류

일본 내시경학회에 따르면, 종양은 6가지 유형으로 분류된다(표 2).<sup>8</sup> 표면형은 T1 종양의 전형적인 형태로, 조기 위암의 육안적 분류에 따라 세분화된다. 진행성 위암(Advanced Gastric Cancer, AGC)은 보만 분류법(Borrmann classification)에 따라 1~4유형으로 분류되고, 이는 그림 7에 자세히 나타나 있다. 조기 위암과 진행성 위암으로 분류할 수 없는 경우에는 5유형에 해당한다.

표 2. 육안적 분류

분류	설명
Type 0 (표면형)	전형적인 T1 종양
Type 1 (용종형)	용종 종양이 주변 점막층과 현저하게 구분되는 경우
Type 2 (궤양형)	가장자리가 올라간 궤양성 종양으로, 가장자리가 뚜렷한 두꺼운 위벽으로 둘러싸여 있는 경우
Type 3 (궤양 침윤형)	가장자리가 올라간 궤양성 종양으로, 가장자리가 뚜렷하지 않은 두꺼운 위벽으로 둘러싸여 있는 경우
Type 4 (미만형)	뚜렷한 궤양이 없거나 가장자리가 올라가지 않은 종양이고, 위벽은 두꺼워지고 단단하며 가장자리가 불분명한 경우
Type 5 (분류 불가)	다른 유형 중 어느 것으로도 종양을 분류할 수 없는 경우

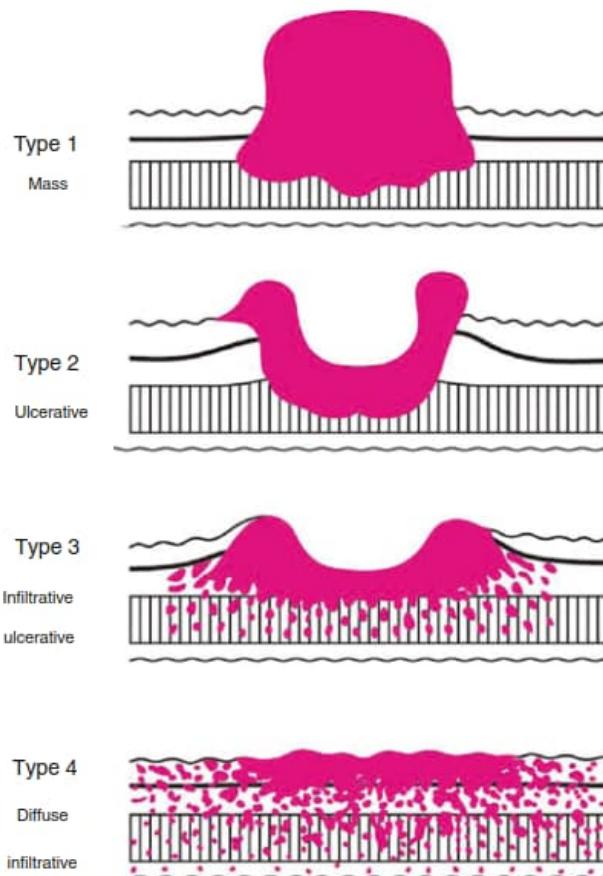


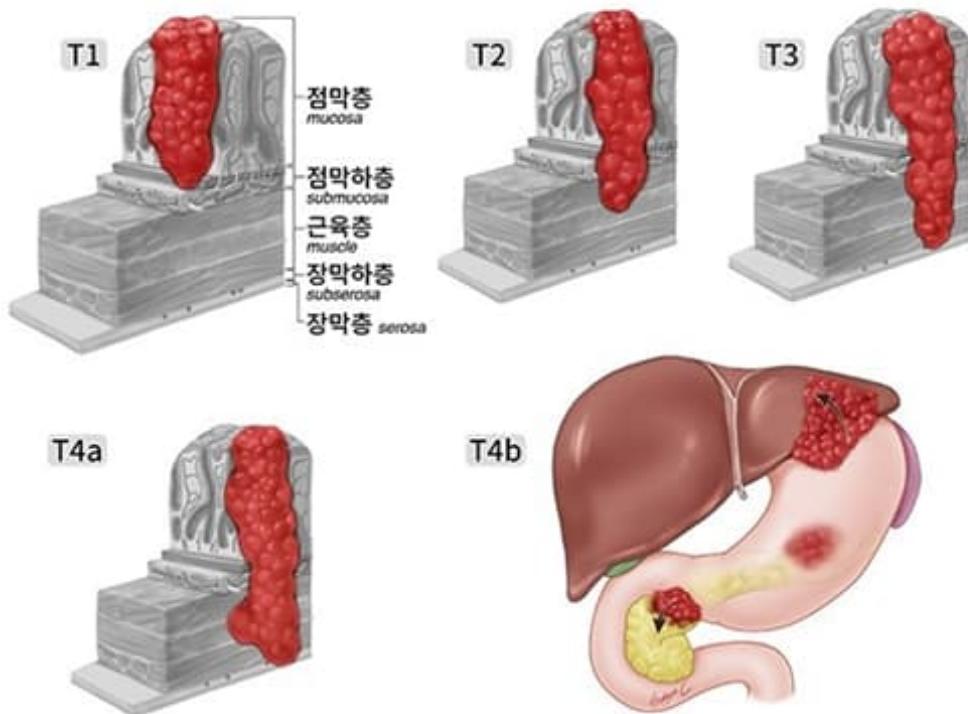
그림 7. 진행성 위암의 육안적 분류

종양의 침습 깊이는 T병기로 구분된다(표 3).<sup>8</sup> 위의 림프절 전이와 상관없이 암 침습이 점막(Mucosa, M) 또는 점막하(Submucosa, SM)층에 국한되어 있는 T1의 경우 조기 위암으로 정의된다.<sup>10</sup> 진행성 위암은 점막하층을 지나 고유근층(Muscularis propria) 이상 침습한 위암으로 T2, T3, T4(T4a, T4b)가 이에 해당한다. 조기 위암 환자의 경우 예후가 우수한 반면, 진행성 위암 환자의 예후는 좋지 않다.<sup>11</sup> 위암 환자의 5년 생존율은 진단 시점의 위암 단계와 높은 상관관계가 있으므로 위암을 조기에 발견하는 것은 중요하다.<sup>4</sup>



표 3. T병기

분류	설명
TX	종양의 침습 깊이 알 수 없는 경우
T0	원발성 종양의 증거가 없는 경우
T1	종양이 점막(Mucosa) 또는 점막하(Submucosa)에 국한된 경우
T1a	종양이 점막(Mucosa)에 국한된 경우
T1b	종양이 점막하(Submucosa)에 국한된 경우
T2	종양이 고유근총(Muscularis propria)을 침습한 경우
T3	종양이 장막밑총(Subserosa)를 침습한 경우
T4	종양이 장막총(Serosa)에 인접하거나 장막총을 관통하거나 주변 장기를 침습한 경우
T4a	장막총에 인접하거나 장막을 관통하여 복막강에 노출된 경우
T4b	종양이 주변 장기를 침습한 경우



감수: 윤충안 (국립암센터 외과)

출처: 국가암정보센터 암정보

그림 8. T병기 도식화

T병기에 따라 조기 위암과 진행성 위암을 진단하는 것은 위암 치료를 위해 중요하다. 하지만 종양의 침습 깊이에 따라 임상 및 병리학적 진단이 때때로 다를 수 있다.<sup>12</sup> 진단 불일치가 발생하는 경우에는 세부적인 임상적 특징이 명확하지 않은 경우가 많아 해결해야 할 중요한 문제로 남아 있다.

## 2. 영상검출·진단보조 소프트웨어의 이해

의료기기 소프트웨어란 의료기기에 해당하는 목적으로 사용하기 위해 개발된 소프트웨어로 독립형 소프트웨어와 내장형 소프트웨어, 모바일 의료용 앱 등을 포함하는 의료기기이다.<sup>13</sup> 이 중 영상검출 및 진단보조 관련 소프트웨어는 어떤 부위를 검출한 후 윤곽선, 색상 또는 지시선 등으로 표시하거나 질병의 유무, 질병의 중증도 또는 질병의 상태 등에 대한 가능성 정도를 자동으로 표시하여 의료인의 진단결정을 보조하는 데 사용하는 소프트웨어로, 국내에는 14개의 품목이 정의되어 있다(표 4).

표 4. 영상검출·진단보조 소프트웨어 품목분류

중분류	품목코드 [등급]	품목명	검출 부위
심혈관 진료용 소프트웨어	E01120.01 [3]	심혈관영상검출·진단 보조소프트웨어	심혈관영상 내에서 정상과 다른 이상 부위
치의학 진료용 소프트웨어	E02030.01 [2]	치과영상검출·진단보 조소프트웨어	치과영상 내에서 정상과 다른 이상 부위
	E04020.01 [2]	내시경영상검출·진단 보조소프트웨어	내시경 영상 내에서 정상과 다른 이상 부위
위장병학 및 비뇨의학	E04020.02 [3]	위암영상검출·진단보 조소프트웨어	내시경영상 내에서 위암 의심부위
진료용 소프트웨어	E04020.03 [3]	대장암영상검출·진단 보조소프트웨어	시경영상 내에서 대장암 의심부위
	E04020.04 [3]	전립선암영상검출·진 단보조소프트웨어	의료영상 내에서 전립선암 의심부위

병원진료용 소프트웨어	E05050.01 [2]	2등급초음파영상검출 · 진단보조소프트웨어	초음파영상 내에서 정상과 다른 이상 부위
	E05050.02 [3]	3등급초음파영상검출 · 진단보조소프트웨어	초음파영상 내에서 암 의심부위
신경과학 진료용 소프트웨어	E06090.01 [3]	뇌영상검출 · 진단보조 소프트웨어	뇌영상 내에서 정상과 다른 이상 부위
안과학 진료용 소프트웨어	E08020.01 [3]	안과영상검출 · 진단보 조소프트웨어	안과영상 내에서 정상과 다른 이상 부위
정형외과학 진료용 소프트웨어	E09020.01 [2]	정형외과영상검출 · 진 단보조소프트웨어	정형외과영상 내에서 정상과 다른 이상 부위
방사선종양학 및 영상의학 진료용 소프트웨어	E11030.01 [2]	2등급의료영상검출 · 진 단보조소프트웨어	의료영상 내에서 정상과 다른 이상 부위
	E11030.02 [3]	3등급의료영상검출 · 진 단보조소프트웨어	의료영상 내에서 암 의심 부위
	E11030.03 [3]	유방암영상검출 · 진단 보조소프트웨어	의료영상 내에서 유방암 의심부위

### 가. 국내 영상검출·진단보조 소프트웨어 허가 현황

국내에서 영상검출·진단보조 의료기기 소프트웨어로 허가된 제품은 제조 및 수입을 합하여 총 46건이다. 위암영상검출·진단보조소프트웨어는 1건이 확인되었으며, 그 외에 2등급의료영상검출·진단보조소프트웨어 15건, 2등급초음파영상검출·진단보조소프트웨어 4건, 3등급의료영상검출·진단보조소프트웨어 1건, 내시경영상검출·진단보조소프트웨어 4건, 뇌영상검출·진단보조소프트웨어 6건, 심혈관영상검출·진단보조소프트웨어 2건, 안과영상검출·진단보조소프트웨어 7건, 유방암영상검출·진단보조소프트웨어 2건, 전립선암영상검출·진단보조소프트웨어 1건, 정형외과영상검출·진단보조소프트웨어 3건을 확인할 수 있었다(표 5). (2022.12 기준)

표 5. 국내 영상검출·진단보조 소프트웨어 허가 현황

번호	품목명 [등급]	업체명	모델명	구분
1	2등급의료영상검출·진단보조소프트웨어 [2]	(주)루닛	Lunit INSIGHT CXR Triage	제조
2	2등급의료영상검출·진단보조소프트웨어 [2]	프로메디우스 주식회사	CXR-02	제조
3	2등급의료영상검출·진단보조소프트웨어 [2]	제이피아이 헬스케어 (주)시화현장	ExamVue Duo AI Plus	제조
4	2등급의료영상검출·진단보조소프트웨어 [2]	메디컬아이피(주)	TiSepX	제조

5	2등급의료영상검출·진단보조소프트웨어 [2]	지멘스헬시니어스(주)	AI-Rad Companion Chest CT	수입
6	2등급의료영상검출·진단보조소프트웨어 [2]	(주)딥노이드	DC-XR-03	제조
7	2등급의료영상검출·진단보조소프트웨어 [2]	(주)엔티엘헬스케어	CerviCare AI	제조
8	2등급의료영상검출·진단보조소프트웨어 [2]	(주)신한씨스텍	MIM Symphony Dx-Lite	수입
9	2등급의료영상검출·진단보조소프트웨어 [2]	(주)루닛	Lunit INSIGHT CXR	제조
10	2등급의료영상검출·진단보조소프트웨어 [2]	(주)코아라인소프트	AVIEW LUNG Nodule CAD	제조
11	2등급의료영상검출·진단보조소프트웨어 [2]	주식회사 뷔노	VN-M-04	제조
12	2등급의료영상검출·진단보조소프트웨어 [2]	(주)레이언스	Xmaru Pro CXR	제조
13	2등급의료영상검출·진단보조소프트웨어 [2]	(주)래디센	AXIR-CX	제조
14	2등급의료영상검출·진단보조소프트웨어 [2]	주식회사 뷔노	VN-M-02	제조
15	2등급의료영상검출·진단보조소프트웨어 [2]	(주)신한씨스텍	BioJet	수입

16	2등급초음파영상검출·진단보조소프트웨어 [2]	(주)필립스코리아	TOMTEC-AR ENA	수입
17	2등급초음파영상검출·진단보조소프트웨어 [2]	모니터코퍼레이션 주식회사	MLA-01	제조
18	2등급초음파영상검출·진단보조소프트웨어 [2]	와이솔루션스	Image-Arena	수입
19	2등급초음파영상검출·진단보조소프트웨어 [2]	와이솔루션스	TomTec-Arena	수입
20	3등급의료영상검출·진단보조소프트웨어 [3]	모니터코퍼레이션 주식회사	ML-02	제조
21	내시경영상검출·진단보조소프트웨어 [2]	(주)엔도아이	endo Ex-1	제조
22	내시경영상검출·진단보조소프트웨어 [2]	광립메디텍	Gi9000-Endo	제조
23	내시경영상검출·진단보조소프트웨어 [2]	(주)엔도아이	endo K-Doc	제조
24	내시경영상검출·진단보조소프트웨어 [2]	(주)인피니트헬스케어	INFINITT Smart Endo	제조
25	뇌영상검출·진단보조소프트웨어 [3]	(주)휴런	ST-AS01, ST-AS02	제조
	뇌영상검출·진단보조소프트웨어 [3]	(주)휴런	ST-HS01	제조
27	뇌영상검출·진단보조소프트웨어 [3]	(주)코어라인소프트	AVIEW NeuroCAD	제조



28	뇌영상검출·진단보조 프트웨어 [3]	SK(주)	SKH-BCH-001	제조
29	뇌영상검출·진단보조 프트웨어 [3]	(주)휴런	H-PD-D01, H-PD-D02, H-PD-D05, mPDia-01	제조
30	뇌영상검출·진단보조 프트웨어 [3]	주식회사 뷰노	VN-M-07	제조
31	심혈관영상검출·진단보 조소프트웨어 [3]	(주)코어라인소프트	AVIEW CAC	제조
32	심혈관영상검출·진단보 조소프트웨어 [3]	(주)새한엔케이엔디	QAngio XA 3D	수입
33	안과영상검출·진단보조 소프트웨어 [3]	(주)에이아이인사이트	Whisky	제조
34	안과영상검출·진단보조 소프트웨어 [3]	(주)에이아이인사이트	Whisky-DR	제조
35	안과영상검출·진단보조 소프트웨어 [3]	(주)에이아이인사이트	Whisky-AMD	제조
36	안과영상검출·진단보조 소프트웨어 [3]	(주)에이아이인사이트	Whisky-GLC	제조
37	안과영상검출·진단보조 소프트웨어 [3]	주식회사 에임즈	EyeView	제조
38	안과영상검출·진단보조 소프트웨어 [3]	(주)메디웨일	DrNoon for Fundus screening	제조



39	안과영상검출·진단보조 소프트웨어 [3]	주식회사 뷰노	VN-M-03	제조
40	위암영상검출·진단보조 소프트웨어 [3]	주식회사 웨이센	WME-01	제조
41	유방암영상검출·진단보 조소프트웨어 [3]	지멘스헬시니어스(주)	Transpara	수입
42	유방암영상검출·진단보 조소프트웨어 [3]	(주)루닛	MMG-WEB, MMG-DCM, Lunit INSIGHT MMG	제조
43	전립선암영상검출·진단 보조소프트웨어 [3]	(주)제이엘케이	JPC-01K	제조
44	정형외과영상검출·진단 보조소프트웨어 [3]	(주)크레스콤	MDAI-FXWR- 01	제조
45	정형외과영상검출·진단 보조소프트웨어 [3]	주식회사 바스젠바이오	DR-SPINE	제조
46	정형외과영상검출·진단 보조소프트웨어 [3]	(주)딥노이드	DS-CF-01	제조

#### 나. 의료기기 소프트웨어 관련 규제

의료기기 소프트웨어의 필수 성능과 안전성을 보장하기 위해 다양한 국제규격이 제정되었다. 의료기기 품질관리, 의료기기 위험관리, 의료기기 소프트웨어 전 주기 프로세스, 의료기기 소프트웨어의 위험관리, 네트워크로 연결되는 의료기기의 위험관리에 대한 내용을 다루고 있다(표 6).

ISO 13485에 따라 품질관리 활동을 수행하고, ISO 14971과 의료기기 소프트웨어에 ISO 14971을 적용하기 위한 가이던스인 IEC/TR 80002-1에 따라 위험관리 활동을 수행해야 한다. 이와 더불어 의료기기 소프트웨어의 전 수명주기 프로세스에 대한 규격인 IEC 62304에 따라 소프트웨어 검증 및 유효성을 확인해야 한다. 소프트웨어의 신뢰도를 높이고, 사용자와 환자에 대한 위험을 낮추어 의료기기로서의 사용목적을 달성할 수 있음을 보여야 한다.

다른 의료기기와 연결하여 사용하거나 유·무선 통신(블루투스, Wi-fi, USB 등)이 가능한 의료기기 소프트웨어는 사이버 공간에서 정보의 기밀성(Confidentiality), 가용성(Availability), 무결성(Integrity)을 보존하기 위해 비인가된 활동으로부터 정보와 시스템을 보호하는 사이버보안 요구사항을 따르는 것도 요구된다.

표 6. 소프트웨어 관련 국외 규격

번호	규격 번호	규격명
1	ISO 13485	Medical devices – Quality management systems – Requirements for regulatory purposes
2	ISO 14971	Medical devices – Application of risk management to medical devices
3	IEC 62304	Medical device software – Software life cycle processes
4	IEC/TR 80002-1	Medical device software – Part 1: Guidance on the application of ISO 14971 to medical device software
5	IEC 80001-1	Application of risk management for IT-networks incorporating medical devices – Part 1: Roles, responsibilities, and activities
6	IEC/TR 80001-2-1	Application of risk management for IT-networks incorporating medical devices – Part 2-1: Step by step risk management of medical IT-networks – Practical applications and examples
7	IEC/TR 80001-2-2	Application of risk management for IT-networks incorporating medical devices – Part 2-2: Guidance for the disclosure and communication of medical device security needs, risks and controls

국내 식품의약품안전처(이하 식약처)에서도 가이드라인을 발간하여 국제규격의 적용을 보조하고 있다. 소프트웨어 및 인공지능 관련 국내 가이드라인을 조사하여 정리하였다(표 7). ‘의료기기 소프트웨어 밸리데이션 가이드라인’에 따라 소프트웨어 검증 및 유효성확인 활동을 수행하고, ‘의료기기 소프트웨어

허가·심사 가이드라인’에 따라 기술문서 작성방법과 제출해야 하는 첨부자료에 대해 도움을 얻을 수 있다.

‘인공지능 기반 의료기기의 임상 유효성 평가 가이드라인’에서는 인공지능 기반 의료기기인 독립형 소프트웨어 의료기기를 대상으로 임상적 유효성 평가와 임상시험 설계 시 고려해야 할 사항들을 제시하고 있다. ‘빅데이터 및 인공지능 기술이 적용된 의료기기 허가·심사 가이드라인’에서는 빅데이터 및 인공지능 기술이 적용된 제품들에 대해 구체적인 허가·심사 방안을 제시하고 있고, ‘인공지능 의료기기의 허가·심사 가이드라인’에서는 기계학습 모델이 적용된 의료기기인 기계학습 가능 의료기기(Machine Learning-enabled Medical Devices, MLMD)를 대상으로 구체적인 허가·심사 방안을 제시하고 있다. MLMD의 임상적 유효성 평가를 위해 ‘인공지능(AI) 의료기기 임상시험방법 설계 가이드라인’에서 임상시험 설계 시 고려 사항을 제시하고 있다.

표 7. 소프트웨어 관련 국내 가이드라인

번호	제·개정 번호	가이드라인명
1	안내서-0590-01	휴대형의료영상전송장치소프트웨어 허가·심사 가이드라인(민원인 안내서)
2	안내서-0592-01	의료영상전송장치소프트웨어 기술문서 작성을 위한 가이드라인(민원인 안내서)
3	안내서-0095-01	의료기기 소프트웨어 뱌리 데이션 가이드라인(민원인 안내서)
4	안내서-0612-03	의료기기 소프트웨어 허가·심사 가이드라인(민원인 안내서)
5	안내서-0590-02	휴대형의료영상전송장치소프트웨어의 기술문서 작성을 위한 가이드라인(민원인 안내서)

6	안내서-1084-01	뇌 영상검출·진단보조 소프트웨어 안전성·성능 및 임상시험계획서 평가 가이드라인(민원인 안내서)
7	안내서-1085-01	전립선암 영상검출·진단보조 소프트웨어 안전성·성능 및 임상시험계획서 평가 가이드라인(민원인 안내서)
8	안내서-1086-01	대장암 영상검출·진단보조 소프트웨어 안전성·성능 및 임상시험계획서 평가 가이드라인(민원인 안내서)
9	안내서-0818-02	인공지능 기반 의료기기의 임상 유효성 평가 가이드라인(민원인 안내서)
10	안내서-0804-02	빅데이터 및 인공지능 기술이 적용된 의료기기 허가심사 가이드라인(민원인 안내서)
11	안내서-0804-03	인공지능 의료기기의 허가·심사 가이드라인(민원인 안내서)
12	안내서-0818-03	인공지능(AI) 의료기기 임상시험방법 설계 가이드라인(민원인 안내서)
13	안내서-0977-02	인공지능 의료기기의 임상시험계획서 작성 가이드라인(민원인 안내서): 관상동맥협착
14	안내서-0978-02	인공지능 의료기기의 임상시험계획서 작성 가이드라인(민원인 안내서): 유방암
15	안내서-0979-02	인공지능 의료기기의 임상시험계획서 작성 가이드라인(민원인 안내서): 허혈성 뇌졸증
16	안내서-0980-02	인공지능 의료기기의 임상시험계획서 작성 가이드라인(민원인 안내서): 폐암/폐결절
17	안내서-0995-02	의료기기의 사이버보안 허가·심사 가이드라인
18	안내서-0996-02	의료기기의 사이버보안 적용방법 및 사례집(민원인 안내서)

#### 다. 영상검출·진단보조 소프트웨어 임상시험의 이해

국내 의료기기 임상시험은 의료기기 임상시험 관리기준(Korea Good Clinical Practice, KGCP)을 준수해야 한다. 품목별로 발간된 식약처 가이드라인이 KGCP에 따라 임상시험을 설계하는 데에 도움이 될 수 있다. 위암영상 검출·진단보조소프트웨어 품목에 대한 가이드라인은 없는 반면, 뇌영상검출·진단보조소프트웨어, 대장암영상검출·진단보조소프트웨어, 전립선암영상검출·진단보조소프트웨어를 대상으로 안전성 및 성능 평가 방법, 임상시험 설계에 대한 가이드라인을 제시하고 있어 프로토콜 개발 시 참고하였다. (2022.12 기준)

임상시험 수행 전에 소프트웨어의 잠재적인 성능을 확인하는 것이 중요하다. 제품의 사용목적에 따라 적절한 성능 평가 항목을 선택할 수 있다. 뇌, 대장암, 전립선암에 대한 영상검출·진단보조 소프트웨어 품목에 대한 가이드라인에서는 성능 평가 항목으로 민감도(Sensitivity), 특이도(Specificity), AUC(Area Under the Curve), ROC(Receiver Operating Characteristic) Curve, 용종 발견율, DSC(Dice Similarity Coefficient)를 제시하고 있으며, 이 중 일부를 제외하거나 추가할 수 있다(표 8).<sup>14-16</sup>

민감도는 특정 질병에 대해 실제 양성인 사람들 중에서 양성으로 분류해 내는 확률이고, 특이도는 실제 음성인 사람들 중에서 음성으로 분류해 내는 확률이다(표 8-9).<sup>17</sup> 진양성(True Positive, TP)은 실제 양성인 사람을 양성으로, 위양성(False Positive, FP)은 실제 음성인 사람을 양성으로, 위음성(False Negative, FN)은 실제 양성인 사람을 음성으로, 진음성(True Negative, TN)은 실제 음성인 사람을 음성으로 분류한 데이터 개수이다(표 9).<sup>18</sup> 양성 예측도(Positive Predictive Value, PPV)는 특정한 특성을 가지고 있는 것으로 분류된 사람들 가운데 실제로 그 특성을 가지고 있는 사람이 차지하는 분율, 음성 예측도(Negative Predictive Value, NPV)는 특정한 특성을 가지고 있지 않

는 것으로 분류된 사람들 가운데 실제로 그 특성을 가지고 있지 않는 사람이 차지하는 분률을 의미한다(표 8-9).<sup>18</sup>

표 8. 성능 평가 항목 정의 및 계산식

평가변수	정의	계산식
민감도	실제로 특정한 질병에 걸린 사람들 중에서 그 질병이 있다고 분류해내는 확률	$\frac{TP}{TP+FN} \times 100\%$
특이도	실제로 특정한 질병이 없는 사람들 중에서 그 질병이 없다고 분류해내는 확률	$\frac{TN}{TN+FP} \times 100\%$
양성 예측도	특정한 특성을 가지고 있는 것으로 분류된 사람들 가운데 실제로 그 특성을 가지고 있는 사람이 차지하는 분률	$\frac{TP}{TP+FP} \times 100\%$
음성 예측도	특정한 특성을 가지고 있지 않는 것으로 분류된 사람들 가운데 실제로 그 특성을 가지고 있지 않는 사람이 차지하는 분률	$\frac{TN}{TN+FN} \times 100\%$
ROC curve	진단검사 결과를 근거로 민감도와 위양성률(1-특이도)을 이용하여 그린 그래프	-
AUC	ROC(Receiver Operating Characteristic) Curve의 아래 면적으로 진단 정확도를 의미	-
용종 발견율	내시경 영상 내 존재하는 모든 용종의 개수와 실제로 찾아낸 용종의 개수의 비율	(찾아낸 용종 수) / (모든 용종 수)
DSC	소프트웨어에서 검출한 병변의 위치와 참조표준에서 결정한 병변의 위치가 얼마나 일치하는지 확인하기 위한 성능지표	$\frac{2 \times  P \cap GT }{ P  +  GT }$ P: Predicted GT: Ground Truth

표 9. 이분형 검사 결과 요약

결과	참조표준		
	양성	음성	전체
양성	진양성(TP)	위양성(FP)	진 양성+위 양성 (TP+FP)
시험군 (소프트웨어)	음성	위음성(FN)	위 음성+진음성 (FN+TN)
전체	진양성+위 음성 (TP+FN)	위 양성+진음성 (FP+TN)	n

ROC curve는 민감도와 위양성률(1-특이도)을 그래프로 나타낸 것으로 병변의 위치를 얼마나 정확하게 검출했는지 알 수 있는 지표이다. AUC는 ROC curve 아래 면적에 해당하며, 0.5에서 1.0 사이의 값에서 1에 근접할수록 이상적인 성능이라고 할 수 있다. 용종 발견율은 내시경 영상 내 존재하는 모든 용종의 개수 대비 실제로 찾은 용종의 비율이고, DSC는 참조표준에서 정한 병변 위치와 소프트웨어가 검출한 병변 위치가 얼마나 일치하는지 확인하는 지표이다. DSC 계산식에서 GT(Ground truth)는 내시경 전문의가 생성한 위암 영역이고, P(Predicted)는 인공지능 모델에서 얻어진 병변의 영역을 의미한다. 성능 평가 항목에 대한 기준은 다양한 선행연구 문헌을 참고하여 설정할 수 있다.<sup>14-16</sup>

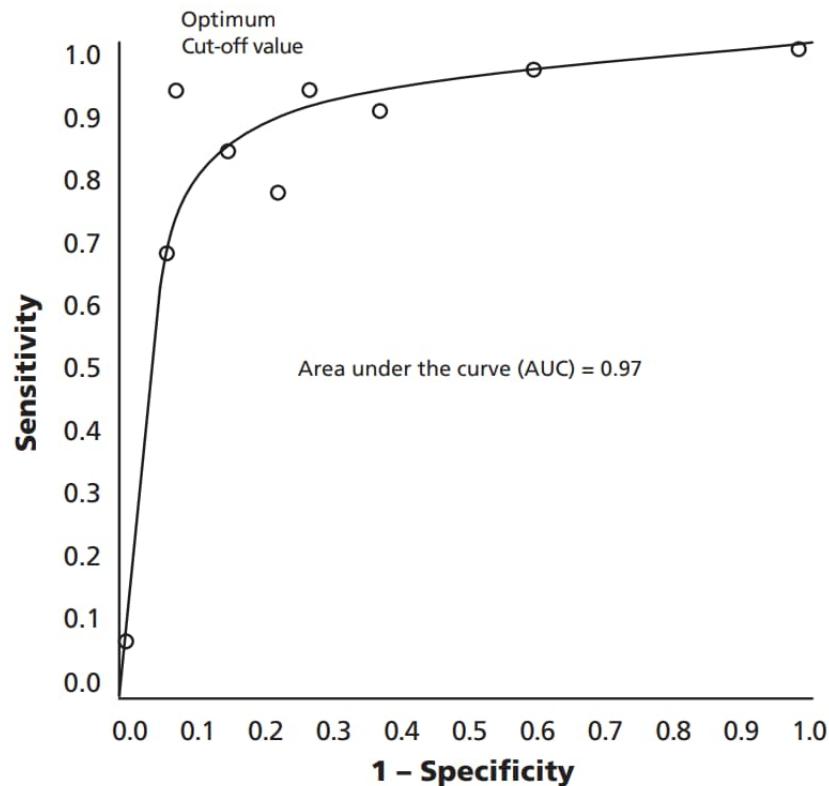


그림 9. ROC curve 및 AUC 그래프<sup>19</sup>

전체 시험 데이터에 대한 평가 결과와 시험 데이터의 특성별 하위 그룹의 성능을 비교한 결과를 모두 시험 결과로 제시 가능하다. 하위그룹은 제품의 성능에 영향을 미칠 수 있다고 예상되는 기준으로 분류하며, 이는 사용목적에 따라 필요하다고 판단되는 경우 적용할 수 있다. 따라서, 본 연구에서 위암(Cancer)과 비위암(Non-Cancer)에 따라 하위 그룹을 분류하였고, 위암 검출 결과의 하위그룹에 대한 성능도 평가항목으로 추가하였다.

### 3. 국내외 위암 인공지능 소프트웨어 성능 평가 및 임상시험 사례

#### 가. 해외 소프트웨어 성능 평가 사례

미국 국립보건원(National Institutes of Health, NIH) 산하 미국 국립의학도서관(National Library of Medicine, NLM)에서 운영하는 의학문헌 검색 시스템인 ‘Pubmed’를 이용하여 위암을 적응증으로 하는 인공지능을 조사하였다. 키워드는 gastric cancer, artificial intelligence를 결합하여 검색하였고, 연관 있는 6개의 사례를 정리하였다(표 10).

표 10. 해외 인공지능 소프트웨어 성능 평가 리스트

번호	연구제목	인공지능 기능
1	Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images	위암 및 비위암 검출
2	Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network	위암 및 비위암 검출
3	Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy	위암의 침습 깊이 진단
4	Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study	위암 및 비위암 검출
5	Automated detection and segmentation of early gastric cancer from endoscopic images using mask R-CNN	EGC 검출
6	Prediction of Submucosal Invasion for Gastric Neoplasms in Endoscopic Images Using Deep-Learning	점막 및 점막하 침습

#### 나. 인공지능 소프트웨어 임상시험 설계 사례

위암을 적응증으로 하고 인공지능 소프트웨어를 대상으로 수행된 임상시험을 조사하고 각 임상시험에 대한 세부사항을 정리하였다. ‘ClinicalTrials.gov’를 이용하여 해외 사례를 조사하고, ‘임상연구정보서비스(Clinical Research Information Service, CRIS)’를 이용하여 국내 사례를 조사하였다.

##### (1) 해외 인공지능 소프트웨어 임상시험 설계 사례

미국 국립의학도서관에서 운영하는 ‘ClinicalTrials.gov’는 임상 연구에 관한 정보를 제공하는 사이트이다. 키워드로는 gastric cancer, gastric disease, artificial intelligence, endoscopy를 조합하여 검색하였고, 검색 결과 중 유사한 연구를 선별 후, 선별하여 조사한 사례를 정리하였다(표 11).

표 11. 해외 인공지능 소프트웨어 임상시험 리스트

번호	NCT 번호	연구제목
1	NCT04040374	A Single-center, Retrospective, Open Label, Randomized Controlled Trial of Artificial Intelligence Versus Expert Endoscopists for Diagnosis of Gastric Cancer in Patients Who Underwent Upper Gastrointestinal Endoscopy
2	NCT04563416	Application of Artificial Intelligence for Early Diagnosis of Gastric Cancer During Optical Enhancement Magnifying Endoscopy

3	NCT03784209	Automatic Real-time Diagnosis of Gastric Mucosal Disease Using Probe-based Confocal Laser Endomicroscopy With Artificial Intelligence
4	NCT04232462	A Multicentric Validation Study on the Accuracy of Artificial Intelligence Assisted System in Clinical Application of Digestive Endoscopy
5	NCT03883035	Utilization of Real-time Automatic Quality-control System in the Detection of Gastric Neoplasms

선별한 임상시험에 대하여 디자인, 선정 및 제외 기준, 유효성 평가변수를 포함한 세부적인 설계 항목들을 정리하였다(표 12-16).

표 12. 해외 인공지능 소프트웨어 임상시험 설계 사례 1

연구 제목	상부위내시경 검사를 받은 환자의 위암 진단을 위한 인공지능과 내시경 전문의의 단일기관, 후향적, 개방적, 무작위 배정 임상시험
적용증	위암
디자인	무작위 배정, 평행설계, 개방적, 중재연구
대상자 수	500 명
선정 기준	<ul style="list-style-type: none"> <li>• 2018년에 도쿄대학병원에서 상부위내시경 검사를 받은 20세 이상의 성인</li> </ul>
제외 기준	<ul style="list-style-type: none"> <li>• 위 절제술을 받은 환자</li> <li>• 비강 상부위내시경 검사를 받은 환자</li> </ul>



---

<b>유효성</b>	<b>일차</b>
<b>평가변수</b>	위암으로 진단받은 대상자 수
<b>유효성</b>	<b>이차</b>
<b>평가변수</b>	위 병변의 IOU(Intersection Over Union)
<b>유효성</b>	<b>일차</b>
<b>평가변수</b>	진행성 위암으로 진단받은 대상자 수
<b>유효성</b>	<b>조기</b>
<b>평가변수</b>	조기 위암으로 진단받은 대상자 수
<b>유효성</b>	<b>인공지능과 내시경 전문의 간 일치</b>
<b>평가변수</b>	인공지능과 내시경 전문의 간 이미지 및 IOU 기반 진단 일치

---

표 13. 해외 인공지능 소프트웨어 임상시험 설계 사례 2

---

<b>연구 제목</b>	광학 강화 확대 내시경검사 중 위암의 조기 진단을 위한 인공지능의 적용
<b>적용증</b>	인공지능, 광학 강화 내시경, 확대 내시경
<b>디자인</b>	관찰연구
<b>대상자 수</b>	80 명
<b>선정 기준</b>	<ul style="list-style-type: none"><li>광학 확대 내시경 검사를 받은 자</li><li>18세 이상</li></ul>
<b>제외 기준</b>	<ul style="list-style-type: none"><li>진행성 위암, 림프종, 활동기 케양, 인공 케양 환자</li></ul>
<b>유효성</b>	<b>일차</b>
<b>평가변수</b>	인공지능의 민감도, 특이도 및 정확도
<b>유효성</b>	<b>이차</b>
<b>평가변수</b>	인공지능의 민감도, 특이도 및 정확도

---



표 14. 해외 인공지능 소프트웨어 임상시험 설계 사례 3

연구 제목	인공지능을 이용한 프로브 기반 공초점 레이저 내시경 검사를 통한 위 점막 질환의 자동 실시간 진단
적용증	위암, 인공지능, 공초점 레이저 내시경
디자인	코호트, 전향적, 관찰연구
대상자 수	951 명
선정 기준	<ul style="list-style-type: none"><li>• 18세 이상 80세 이하</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 응고장애, 신장 또는 간 기능 장애, 임신 또는 모유 수유, 플루오레세인나트륨(fluorescein sodium) 알레르기 등 공초점 레이저 내시경 수행에 부적합 상태의 환자</li></ul>
유효성 평가변수	일차 유효성 평가변수 인공지능의 위 점막 질환 진단 정확도, 민감도, 특이도, PPV, NPV
평가변수	이차 유효성 평가변수 인공지능과 내시경 의사의 위 점막 질환 진단 효율 (정확도, 민감도, 특이도, PPV, NPV) 비교



표 15. 해외 인공지능 소프트웨어 임상시험 설계 사례 4

연구 제목	소화기 내시경의 임상 적용에서 인공지능 보조 시스템의 정확성에 관한 다중심적 검증 연구
적용증	위장병, 내시경, 인공지능
디자인	관찰연구
대상자 수	10,000 명
선정 기준	<ul style="list-style-type: none"><li>• 18세 이상</li><li>• 소화기 질환 특성을 명확히 이해하기 위해 내시경 검사 및 관련 검사를 수행한 자</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 지난 5년간 약물 또는 알코올 남용 또는 심리적 장애를 겪은 자</li><li>• 임산부 또는 수유부</li><li>• 위장 수술 이력이 있는 자</li></ul>
일차 유효성 평가변수	인공지능의 정확도, 민감도, 특이도, PPV, NPV, ROC, AUC
유효성 평가변수	<ul style="list-style-type: none"><li>• 평균 정밀도(mAP, mean Average Precision)</li></ul>
이차 유효성 평가변수	<ul style="list-style-type: none"><li>• Sørensen-Dice 계수</li><li>• 리콜 비율</li></ul>
양성 가능성 비율 평가변수	<ul style="list-style-type: none"><li>• 양성 가능성 비율</li><li>• 음성 가능성 비율</li></ul>



표 16. 해외 인공지능 소프트웨어 임상시험 설계 사례 5

연구 제목	위 종양 검출에 실시간 자동 품질관리 시스템 활용
적용증	위내시경
디자인	무작위 배정, 평행설계, 이중 눈가림, 중재연구
대상자 수	1060 명
선정 기준	<ul style="list-style-type: none"><li>• 40세 이상 80세 이하</li><li>• 조기 위암 검사가 불가능한 자</li><li>• 1년 이내에 위내시경 검사를 받은 자</li><li>• 내시경 검사 결과 상부 위장 악성 병변이 있는 자</li><li>• 상부 위장관 암 병력이 있는 자</li><li>• 마취에 알레르기가 있는 자</li><li>• 협착, 막힘, 고형 음식, 마취 합병증으로 위내시경 수행이 불가능한 자</li></ul>
제외 기준	
일차 유효성 평가변수	위 종양 검출율
유효성 평가변수	<ul style="list-style-type: none"><li>• 검출된 위 종양의 평균 수</li><li>• 절차별 검사 완료도</li><li>• 절차별 검사 시간</li><li>• EAS 지원 집단의 EAS 오류</li></ul>

## (2) 국내 인공지능 소프트웨어 임상시험 설계 사례

식약처에서 제공하는 임상시험승인 현황에서는 임상시험 승인 일자, 의료기기 품목명, 임상시험의 제목만 확인 가능하며, 임상시험 설계에 대한 세부사항은 알 수 없으므로, ‘임상연구정보서비스’를 이용하여 국내 임상시험 설계 사례를 조사하였다. 보건복지부의 지원을 받아 질병관리본부에서 운영하는 ‘임상연구정보서비스’는 국내에서 진행되는 임상시험 및 임상연구를 등록하는 시스템이다. 인공지능을 키워드로 검색한 결과, 적응증을 위암으로 하는 사례는 없었으므로 위내시경과 유사하게 대장내시경과 의료 영상 데이터를 분석한 임상시험을 선별하였다(표 17).

표 17. 국내 인공지능 소프트웨어 임상시험 리스트

번호	등록번호	연구제목
1	KCT0005619	대장내시경 시행 환자를 대상으로 인공지능 Smart Endo의 선종발견율의 우월성을 평가하기 위한 전향적, 다기관, 대조군, 무작위 배정, 공개, 비교 임상시험
2	KCT0005614	인공지능(딥러닝) 기술을 이용한 피부암 진단의 전향적 연구
3	KCT0005591	소장 출혈이 의심되는 환자에서 국산 양방향 MiroCam MC2000 캡슐내시경과 글로벌제품인 단일방향 PillCam SB3 캡슐내시경의 전향 무작위 다기관 비교 연구 : 캡슐인공지능영상연구회 연구
4	KCT0005459	인공지능을 사용한 폐CT영상분석 소프트웨어의 안정성 및 유효성 연구

5	KCT0005065	흉부 CT 영상을 이용한 인공지능 기반 폐 결절 탐지 소프트웨어 VUNO Med - Lung CAD의 임상적 유효성을 평가하기 위한 다기관, 후향적, 확증 임상시험
6	KCT0005051	건강 검진 흉부단순촬영 검사에서의 폐결절 및 폐암 진단: 인공지능 융합형 차세대 PACS의 유효성 검증을 위한 비교 임상시험
7	KCT0005007	응급실 내 급성 흉부 질환 의심 환자에서 흉부 X선 검사의 진단 민감도 평가: 인공지능 기반 컴퓨터 보조 검출 시스템의 유효성 검증을 위한 무작위 비교 임상시험
8	KCT0004902	인공지능 기반 안저영상 판독보조 소프트웨어(VUNO Med - Fundus AI)의 임상적 유효성을 평가하기 위한 단일기관, 단일군, 후향적, 확증적 임상시험
9	KCT0004758	뇌 T1 weighted MR 영상을 이용하여 인공지능 기반 의료영상진단보조소프트웨어의 알츠하이머병 진단 보조 유효성을 평가하기 위한 단일기관, 코호트 내 환자-대조군, 확증 임상시험

선별한 임상시험에 대하여 적응증, 디자인, 선정 및 제외 기준, 유효성 평가변수를 포함한 세부적인 설계 항목들을 정리하였다(표 18-26).



표 18. 국내 인공지능 소프트웨어 임상시험 설계 사례 1

연구 제목	대장내시경 시행 환자를 대상으로 인공지능 Smart Endo의 선종발견율의 우월성을 평가하기 위한 전향적, 다기관, 대조군, 무작위 배정, 공개, 비교 임상시험
적용증	대장내시경
디자인	평행설계, 단일 눈가림, 무작위 배정, 중재연구
대상자 수	1,098 명
선정 기준	<ul style="list-style-type: none"><li>• 만 20세 이상 80세 이하의 성인</li><li>• 대장 절제술을 받은 환자</li><li>• 염증성 장질환 환자</li><li>• Colonic polyposis syndrome 환자</li><li>• 연구 참여를 거부한 환자</li><li>• Dual antiplatelet 제제를 복용 중인 환자 중 약제 중단이 불가하여 조직검사나 용종절제가 불가능한 환자</li><li>• 대장내시경 검사의 절대적 금기에 해당하는 자</li><li>• 이전에 대장내시경에 실패한 자</li><li>• Bowel obstruction이나 perforation 이 있거나 의심되는 자</li><li>• 현재 임신 중이거나 수유 중인 자</li><li>• 치료 시술을 위한 내시경을 시행하는 환자 (예. 용종제거를 위한 내시경)</li></ul>
제외 기준	
일차 유효성 평가변수	군 별 대장선종발견율 비교
유효성 평가변수	<ul style="list-style-type: none"><li>• 군 별 대장 용종 발견율 비교</li></ul>
이차 유효성 평가변수	<ul style="list-style-type: none"><li>• 군 별 withdrawal time(회수 시간) 비교</li><li>• 3년이상의 내시경 전문가와 인공지능과의 선종발견율 비교</li><li>• 초심자와 인공지능과의 선종발견율 비교</li></ul>



표 19. 국내 인공지능 소프트웨어 임상시험 설계 사례 2

연구 제목	인공지능(딥러닝) 기술을 이용한 피부암 진단의 전향적 연구
적응증	피부 병변
디자인	평행설계, 무작위 배정, 중재연구
대상자 수	600 명
선정 기준	<ul style="list-style-type: none"><li>• 19세 이상</li><li>• 환자 또는 의사가 피부암을 의심하는 케이스</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 피부암 생검을 이미 시행한 자</li></ul>
유효성	일차 평가변수
평가변수	이차 유효성    피부종양 진단의 정확도 평가변수



표 20. 국내 인공지능 소프트웨어 임상시험 설계 사례 3

연구 제목	소장 출혈이 의심되는 환자에서 국산 양방향 MiroCam MC2000 캡슐내시경과 글로벌제품인 단일 방향 PillCam SB3 캡슐내시경의 전향 무작위 다기관 비교 연구 : 캡슐인공지능영상연구회 연구
적응증	소장 출혈
디자인	평행설계, 개방적, 무작위배정, 중재연구
대상자 수	172 명
선정 기준	<ul style="list-style-type: none"><li>• 19세 이상 85세 이하</li><li>• 소장 출혈이 의심되는 환자로, 1)반복적 또는 지속적인 철 결핍성 빈혈이 있거나 대변잠혈검사 양성 또는 실제 눈에 보이는 출혈이 있어 진료 시 캡슐 내시경 검사를 시행하기로 결정된 환자 혹은 2)캡슐내시경 시행 전 최소 6개월 이내의 위내시경과 대장내시경 검사에서 뚜렷한 출혈의 원인이 발견되지 않아 진료 시 캡슐내시경 검사를 시행하기로 결정된 환자</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 혈역동학적으로 불안정한 경우</li><li>• 이전에 캡슐내시경으로 전 소장 검사 진행이 안된 경우</li><li>• 위장관 마비, 샛길, 협착이 의심되는 경우</li><li>• 삼킴 장애가 있는 경우</li><li>• Zenker계실의 기왕력이 있는 경우</li><li>• 이전에 소장절제 수술력이 있는 환자</li><li>• 유전성 위장관 용종증이 있는 환자</li><li>• 염증성장질환이 있는 환자</li><li>• 말기 신부전 또는 심부전 환자로 장정결제 복용이 어려운 경우</li><li>• 임산부</li></ul>



- 
- 신경정신학적 질환이 있는 환자
  - 피험자의 자발적 동의가 불가능한 경우
  - 인공 심장 보조기(예, 삽입형 재세동기, 인공 심박 동기, 심실보조장치 등)를 가지고 있는 환자는 절대 배제기준은 아니지만, 환자의 컨디션에 따라 장정결제 복용과 검사 진행에 어려움이 있을 것으로 판단되는 경우

---

유효성 평가변수	일차 유효성 평가변수	MiroCam MC2000과 PillCam SB3의 출혈의 원인 진단율
	이차 유효성 평가변수	MiroCam MC2000과 PillCam SB3의 바тер씨 팽대부 발견율

---

표 21. 국내 인공지능 소프트웨어 임상시험 설계 사례 4

---

연구 제목	인공지능을 사용한 폐CT영상분석 소프트웨어의 안정성 및 유효성 연구
적응증	만성 폐쇄성 폐질환(COPD), 폐암
디자인	단일군, 개방적, 중재연구
대상자 수	200 명
선정 기준	<ul style="list-style-type: none"><li>• 20세 이상</li><li>• 만성 폐쇄성 폐질환 확진 환자</li><li>• 흡연력이 있는 폐암 의심 환자</li><li>• 조기 COPD 의심 환자</li></ul>

---



---

<b>제외 기준</b>	<ul style="list-style-type: none"><li>폐렴, 결핵, 기흉 및 흉막삼출액을 동반한 폐질환 환자</li><li>충분한 흡기를 하고 시행하지 못하거나 촬영 중 움직임이 있는 흉부 CT 영상</li><li>폐 수술을 받은 환자</li><li>한 개의 폐엽 이상을 침습한 과거 염증의 후유증이 있는 영상</li></ul>
<b>일차 유효성 평가변수</b>	후향적 데이터인 Percentile rank 분석으로 분류한 대상 중 COPD 환자, 조기 COPD 의심환자, 폐암 의심 환자 200명을 선정, 전향적으로 모집한 대상에게 소프트웨어를 사용한 지표를 보여준 200명과의 금연율을 독립표본 t-test로 두 그룹간 차이를 검정
<b>유효성 평가변수</b>	전향적으로 모집한 200명을 추적관찰을 통하여 급성 악화, 호흡곤란 등 임상지표를 활용하여 악화군과 안정군 두 그룹으로 분류하고 각 그룹에서 소프트웨어로 얻은 CT 지표들을 독립표본 t-test로 두 그룹간 차이를 검정

---



표 22. 국내 인공지능 소프트웨어 임상시험 설계 사례 5

연구 제목	흉부 CT 영상을 이용한 인공지능 기반 폐 결절 탐지 소프트웨어 VUNO Med - Lung CAD의 임상적 유 효성을 평가하기 위한 다기관, 후향적, 확증 임상시험
적용증	폐 결절
디자인	단일군, 개방적, 중재연구
대상자 수	855 명
선정 기준	<ul style="list-style-type: none"><li>• 2012년 1월부터 2018년 6월까지 흉부 CT 검사를 받은 만 19세 이상의 성인</li><li>• 흉부 CT 검사 결과 폐 결절이 없거나 크기가 4 mm 이상 30 mm 이하(장축 기준)인 1개 이상 5개 이하의 결절이 확인된 자</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 흉부 CT 촬영일 기준 임신 또는 수유중인 여성</li><li>• 흉부 CT 촬영일 이전 5년 이내 폐암 이외의 암 진단을 받은 자</li><li>• 흉부 CT 촬영일 이전 1개월 이내 다음 중 하나 이상의 병력이 있는 자<ul style="list-style-type: none"><li>- 심한 폐섬유화증</li><li>- 미만성 기관지확장증</li><li>- 광범위한 염증성 폐경화</li><li>- 대량의 흉막 삼출</li><li>- 활성 또는 잠복 결핵</li></ul></li><li>• 흉부 CT 영상이 아래 중 하나 이상에 해당되어 적절한 판독이 어려운 경우<ul style="list-style-type: none"><li>- 검사 부위가 불충분하게 포함되어 있는 경우 (예: 양측 폐가 완전히 보이지 않는 경우, 폐첨 부보다 적어도 한 장 이상과 양측 부신이 포함 되어 있지 않은 경우)</li><li>- 환자 움직임 또는 장비 문제로 인한 심한 허상</li></ul></li></ul>



---

(motion artifact)

- 절편 두께가 5 mm를 초과하는 경우
  - 절편 간격(interslice gap)이 있는 경우
  - 표준 재구성 알고리즘(reconstruction kernel)이 아닌 경우
  - 해상도가 낮아 폐와 종격동 구조물의 평가가 어려운 경우
- 

일차

유효성 병변 기반 민감도

평가 변수

이차

유효성 • 피험자 기반 민감도, 특이도, 위양성도, 위음성도

평가 변수

- 병변 기반 위음성도
-



표 23. 국내 인공지능 소프트웨어 임상시험 설계 사례 6

연구 제목	건강 검진 흉부단순촬영 검사에서의 폐결절 및 폐암 진단: 인공지능 융합형 차세대 PACS의 유효성 검증을 위한 비교 임상시험
적용증	폐 결절
디자인	평행설계, 개방적, 무작위 배정, 중재연구
대상자 수	84,000 명
선정 기준	<ul style="list-style-type: none"><li>• 19세 이상</li><li>• 연구기간 동안 서울대학교병원 건강검진센터에서 흉부단순촬영을 시행한 모든 19세 이상 성인</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 흉부단순촬영 시행시 호흡기 증상이 있는 자</li></ul>
유효성 평가변수	<p>일차 유효성 평가변수 폐결절 진단율</p> <p>이차 유효성 평가변수 전체 양성율, 민감도, 특이도, 양성예측도, 음성예측도</p> <p>폐암진단율, 기타 폐질환 진단율 등</p>



표 24. 국내 인공지능 소프트웨어 임상시험 설계 사례 7

연구 제목	응급실 내 급성 흉부 질환 의심 환자에서 흉부 X선 검사의 진단 민감도 평가: 인공지능 기반 컴퓨터 보조 검출 시스템의 유효성 검증을 위한 무작위 비교 임상시험
적응증	폐렴, 폐결핵, 폐부종, 흉막 삼출, 기흉, 폐종양
디자인	평행설계, 개방적, 무작위 배정, 중재연구
대상자 수	4,862 명
선정 기준	<ul style="list-style-type: none"><li>연구 기간 동안 서울대학교병원 응급실을 방문한 만 19세 이상의 성인 환자</li><li>응급실 방문 당시 주증상이 오한, 기침, 호흡곤란, 발열, 객혈, 흉통, 객담 중 어느 하나에 해당하는 환자</li><li>서울대학교병원에서 흉부 X선 검사를 촬영한 환자</li></ul>
제외 기준	<ul style="list-style-type: none"><li>한국 응급환자 중증도 분류기준 (KTAS) 1등급에 해당하는 중증 환자</li><li>외상으로 응급실을 내원한 환자</li></ul>
유효성 평가변수	<b>일차</b> 유효성 급성 흉부 질환에 대한 흉부 X선 판독의 민감도 <b>평가변수</b>
평가변수	<b>이차</b> 유효성 급성 흉부 질환에 대한 흉부 X선 판독의 위양성을 등 <b>평가변수</b>



표 25. 국내 인공지능 소프트웨어 임상시험 설계 사례 8

연구 제목	인공지능 기반 안저영상 판독보조 소프트웨어(VUNO Med - Fundus AI)의 임상적 유효성을 평가하기 위한 단일기관, 단일군, 후향적, 확증적 임상시험
적용증	망막안저영상검사
디자인	단일군, 개방적, 중재연구
대상자 수	1,713 명
선정 기준	<ul style="list-style-type: none"><li>• 2016년 7월부터 2018년 6월까지 안저검사를 받은 만 19세 이상의 성인</li><li>• 건강검진 결과문 또는 의무기록상 진단명 등을 통해 안저영상의 이상소견 유무를 확인할 수 있는 경우</li></ul>
제외 기준	<ul style="list-style-type: none"><li>• 안저사진을 촬영한 안구에 심한 백내장, 각막 혼탁 등 매체 혼탁이 있어 안저영상 판독이 적합하지 않은 경우</li></ul>
유효성 평가변수	일차 유효성 평가변수 12개 소견에 대한 소견별 곡선하면적: 수신자조작특성곡선의 아래 면적
평가변수	이차 유효성 평가변수 <ul style="list-style-type: none"><li>• 이상소견별 민감도, 특이도</li><li>• 전반적 정확도, 민감도, 특이도</li></ul>



표 26. 국내 인공지능 소프트웨어 임상시험 설계 사례 9

연구 제목	뇌 T1 weighted MR 영상을 이용하여 인공지능 기반 의료영상진단보조소프트웨어의 알츠하이머병 진단 보조 유효성을 평가하기 위한 단일기관, 코호트 내 환자-대조군, 확증 임상시험
적응증	알츠하이머병
디자인	단일군, 단일 눈가림, 중재연구
대상자 수	350 명
선정 기준	<ul style="list-style-type: none"><li>• 2010년 1월에서 2019년 9월 사이에 내원하여 뇌 자기공명영상(T1 weighted MRI)을 촬영한 만 50세 이상의 성인</li><li>• 아래 기준에 따라 질환군 또는 정상군으로 분류되는 자<ul style="list-style-type: none"><li>- 질환군: Amyloid PET 검사 결과가 양성인 자로, NINCDS-ADRDA 진단기준에 따라 1)유력 알츠하이머병(Probable AD) 혹은 2)가능 알츠하이머병(Possible AD)으로 진단되거나, International Working Group on Mild Cognitive Impairment 기준에 따라 경도인지장애(MCI)로 진단된 자</li><li>- 정상군: Amyloid PET 검사 결과가 음성인 자로, 1)주관적 인지기능 저하를 호소하지 않는 자 또는 2)객관적인 인지기능 저하가 없는 자</li></ul></li></ul>



### 제외 기준

- 알츠하이머병이 아닌 기타 원인에 의한 치매가 있는 환자
- 뇌 T1 강조 MR 영상에서 인지기능 결손을 초래 할 수 있는 대뇌 병변(예: 공간점유 병변, 경막하 병변 혹은 정상뇌압 수두증 등)이 확인되거나 T2 FLAIR MR 영상에서 백질고강도 신호가 grade 3 이상으로 심한 경우
- 진단일 또는 임상평가일과 MRI 촬영일 사이의 간격이 1년을 초과하는 경우(정상군 및 질환군 중 경도인지장애 환자에 한함)

	일차
유효성	민감도, 특이도
평가 변수	
	이차
유효성	ROC curve의 아래 면적
평가 변수	

#### 4. 임상시험 대상 의료기기 개요

임상시험에 사용되는 대상 의료기기는 위내시경을 통해 획득한 영상을 분석하고 위암 의심 영역을 표시해 주어 위암을 검출하고 의료진의 진단결정을 보조하는 의료기기 소프트웨어다. 딥러닝 기반의 콘볼루션 신경망(Convolutional Neural Network, CNN)을 사용하여 학습된 데이터를 바탕으로 영상의 특징을 추출한다. 내시경 장비의 영상을 실시간으로 연동하여 화면에 출력한다. 인공지능이 특정 병변을 탐지했을 경우 병변의 위치를 표시해 주며, 병변 유형을 분석하여 예측 확률을 함께 표시한다(그림 10).

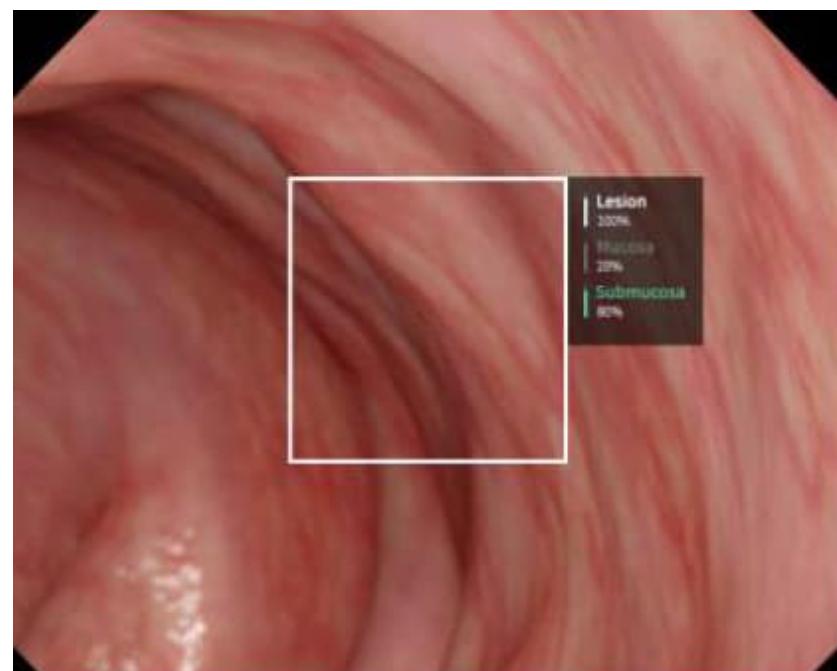


그림 10. 위내시경 영상 분석 화면

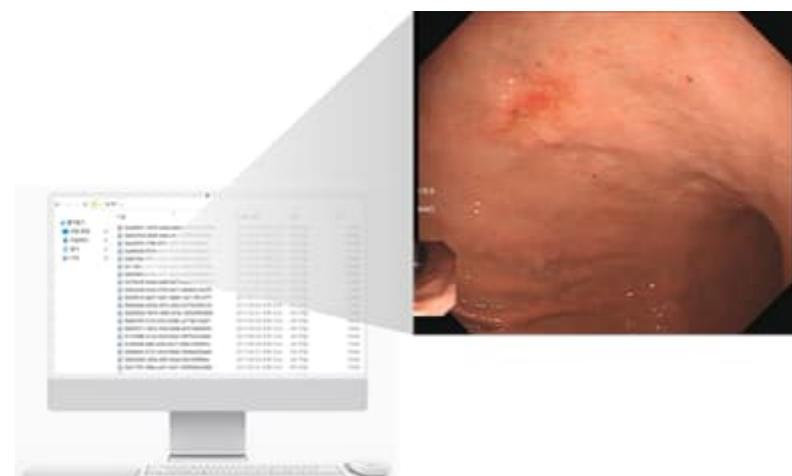
약 15,000개의 위내시경 데이터가 인공지능의 학습에 사용되었으며, 위암의 심 영역에 대한 패턴을 감지하고 표시한 결과 위암과 비위암에 대한 민감도 및 특이도 모두 각각 0.8 이상의 성능을 보였다.

해당 소프트웨어는 크게 메인 모듈, 영상 처리 모듈, 분석 결과 모듈, 데이터 저장 모듈로 구성되어 있다(그림 11). 각 모듈은 소프트웨어의 주요 기능을 수행하는 단위이다. 메인 모듈은 각 모듈 간 통신을 연동해주는 역할을 한다. 영상 처리 모듈은 획득한 내시경 영상을 연산 처리하고 화면에 디스플레이 한다. 분석 결과 모듈은 내시경 검사가 종료 되면 분석 결과를 연산화하는 역할을 한다. 데이터 저장 모듈은 분석 결과를 파일로 저장하고, 저장된 데이터는 데이터베이스(Database)에 쌓이게 된다.



그림 11. 소프트웨어 구성

PC에 소프트웨어를 설치한 후 표본데이터가 저장되어 있는 PC에서 기록대상 데이터를 열면 소프트웨어는 이를 연동하여 실시간으로 병변의 위치를 찾고 예측 결과를 출력함으로써 내시경 전문의의 내시경 검사를 지원한다.



[표본데이터가 보관된 PC]



[표본데이터가 보관된 PC]

[의료기기]

그림 12. PC와 소프트웨어 연동 화면

### III. 결과

#### 1. 위내시경 영상 분석 인공지능 소프트웨어 임상시험 방법

##### 가. 임상시험 설계

앞서 조사한 국내 가이드라인과 국내외 임상시험 설계 사례를 토대로 임상시험을 설계하였다. 임상시험 방법에 따른 피험자 수 산출, 유효성 평가변수 및 평가 방법 관련하여 근거를 제시하는 것이 중요하기 때문에, 프로토콜을 개발하는 과정에서 임상시험의 디자인은 수차례 수정 및 보완 작업을 거쳤다.

‘ClinicalTrials.gov’ 및 ‘임상연구정보서비스’를 이용하여 조사한 인공지능 소프트웨어 임상시험 설계 사례를 보면 인공지능 소프트웨어의 유효성 검증을 위해 대부분 시험군과 대조군의 결과를 비교하는 비교 임상시험으로 설계하였다. 소프트웨어가 의료진의 진단을 보조해 주는 역할이기 때문에 인공지능 단독으로는 하나의 군이 될 수 없다고 판단하여 위암영상검출·진단보조소프트웨어의 지원을 받은 내시경 전문의를 시험군으로 설정하고, 위암영상검출·진단보조소프트웨어의 지원을 받지 않은 내시경 전문의를 대조군으로 설정하였다.

시험군과 대조군의 배치 방법에 따라 평행설계(Parallel design), 교차설계(Crossover design) 등으로 설계할 수 있다. 평행설계는 연구 대상자를 무작위 배정하여 서로 다른 군에 배치하고 연구 종료 시까지 처음 배정된 군에 속하며 다른 군으로 배정되지 않는 반면, 교차설계는 동일한 연구 대상자에게 시간 차이를 두고 시험군과 대조군을 모두 적용한다(표 27).<sup>20</sup> 교차설계는 주로 안전성 및 유효성 정보를 수집하는 탐색 임상시험에서 많이 사용되므로, 확증 비교 임상시험에서 가장 일반적으로 사용되는 평행설계 방법을 사용하였다.

표 27. 평행설계 및 교차설계

	평행설계	교차설계
과정	모집단→스크리닝→무작위배정 →시험군(A)/대조군(B)→평가	모집단→스크리닝→무작위배정 →시험군(A)/대조군(B)→평가 →대조군(A)/시험군(B)→평가
장점	- 직관적 - 단순한 배치 체계 - 명확한 해석 가능	- 자기 대조군 - 상대적으로 적은 표본 수
단점	- 탈락, 결측자료 등 발생 가능 - 상대적으로 많은 표본 필요	- 잔류효과 발생 가능 - 비뚤림 발생 가능

유효성을 입증하기 위해서는 임상시험의 목적에 따라 우월성(Superiority), 비열등성(Non-Inferiority), 동등성(Equivalence) 등의 방법을 고려해야 한다. 우월성 검정은 시험군(임상시험 대상 의료기기)의 효과가 대조군(삼기기 또는 활성 대조군)보다 뛰어남을 보이는 것이고, 비열등성 검정은 시험군의 효과가 대조군보다 열등하지 않음을 보이고, 동등성 검정은 시험군의 효과가 대조군과 유사함을 보이는 것을 목적으로 한다.<sup>20</sup> 우월성 검정에서는 시험군과 대조군의 결과 차이값에 대한 양측 신뢰구간이 0보다 클 때 시험군이 대조군보다 뛰어나다고 할 수 있다(그림 13).<sup>21</sup> 비열등성 검정에서는 비열등성 한계점(margin,  $\Delta$ )을 설정하고, 시험군과 대조군의 결과 차이값에 대한 양측 신뢰구간이 비열등성 한계점( $-\Delta$ )보다 클 때 시험군이 대조군보다 열등하지 않음이 입증된다(그림 14).<sup>21</sup> 동등성 검정은 동등성 한계점( $\Delta$ )을 설정하고, 시험군과 대조군의 결과의 차이값에 대한 양측 신뢰구간이 ( $-\Delta$ ) ~ ( $+\Delta$ ) 구간 내에 있을 때 시험군과 대조군에 유의한 차이가 없음을 입증할 수 있다(그림 15).<sup>21</sup>

인공지능 소프트웨어의 지원을 받는 내시경 전문의가 지원을 받지 않는 내시경 전문의보다 우월함을 보일 수 있도록 우월성 시험으로 설계하였다.

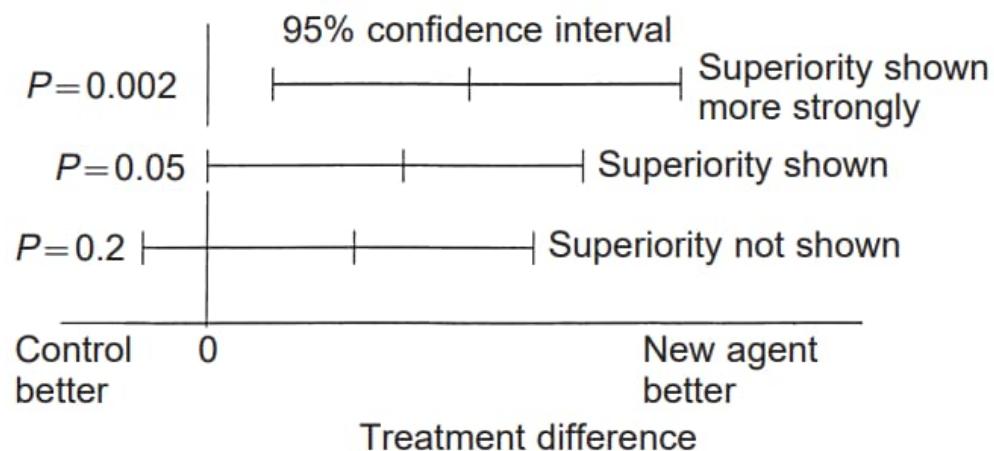


그림 13. 우월성 검정 신뢰구간

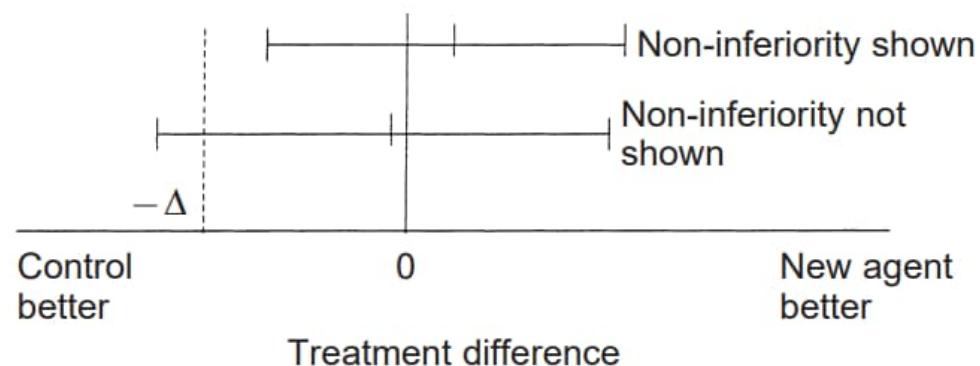


그림 14. 비열등성 검정 신뢰구간

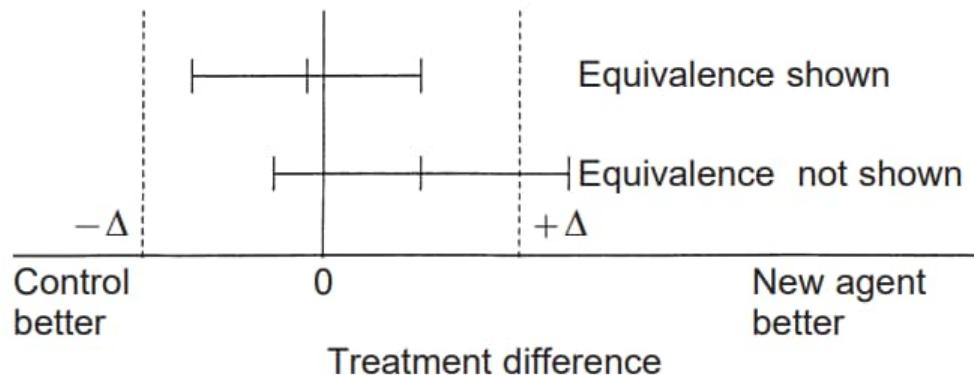


그림 15. 동등성 검정 신뢰구간

표 28. 대조군 종류

sham(sham)기기 대조군	처치의 효과가 없는 기기 사용
무치료 대조군	처치의 효과가 없을 것으로 예상되는 기기 사용
활성 대조군	이미 사용되고 있는 표준기기 또는 처치를 사용
과거 대조군	이전의 임상시험의 자료나 결과를 적용
자기 대조군	동일 대상자에게 시험기기 또는 대조기기를 적용하고 일정한 시차를 두고 대조기기 또는 시험기기를 적용
시차 대조군	같은 조건 하에 시간 또는 장소의 차이를 두고 적용

이를 토대로 일차적으로 대조군, 평행설계, 우월성, 확증 임상시험으로 설계하였고, 임상시험 대상 의료기기의 적용 대상이 되는 대상자가 대상자의 의무기록(Electronic Medical Record, EMR)을 수집한 표본 데이터로 대체되므로 후향적 임상시험으로 설계하였다.

이후, 소프트웨어만을 시험군으로 설정하여도 진단을 보조하는 것으로 간주

한다는 식약처의 의견에 따라 시험군 및 대조군을 수정하였다. 두 군의 결과를 비교하는 대조군 비교 검정 방법은 동일하다. 기존의 시험군은 위암영상검출·진단보조소프트웨어의 지원을 받은 내시경 전문의, 대조군은 위암영상검출·진단보조소프트웨어의 지원을 받지 않은 내시경 전문의로 두 군에 모두 의료진이 포함되어 있어 의료진이 2배로 판독하게 되고, 의료진의 피로도가 상승하게 된다. 따라서, 의료진의 피로도를 상대적으로 감소시킬 수 있도록 위암영상검출·진단보조소프트웨어와 의료진을 각각 시험군과 대조군으로 변경하였다. 식약처 의견에 따르면 소프트웨어는 단독으로 사용할 수 없고 의료진을 보조하는 역할임에도 소프트웨어만을 시험군으로 설정하는 것에 문제가 없다.

또한, 표본 데이터 구축과 평가변수에 관한 설계 사항도 변경하였다. 초기 임상시험 설계 시 소프트웨어가 위내시경 영상을 분석할 때 위암의 유무뿐만 아니라 침습 깊이에 대해 점막(M), 점막하(SM)로 분류할 수 있도록 M, SM 을 포함하는 초기 위암과 진행성 위암도 구분하여 표본 데이터를 구성하고자 하였다. 침습 깊이는 이차적인 요소이므로 표본 데이터 구성 시 고려하지 않아도 된다는 통계적 의견에 따라 임상시험기관의 실제 임상 데이터 수와 통계적인 산출을 통해 EGC:AGC의 비율을 동일하게 구성하였다. 즉, M:SM:AGC 의 비율을 1:1:2로 설정하였다. 제품 허가 시 일차 유효성 평가변수에 해당하는 기능으로만 허가를 받을 수 있기 때문에, 유효성 평가변수로 위암 유무를 일차, 침습 깊이를 이차적인 요소 설정하게 되면 이차로 설정한 침습 깊이에 대한 허가를 받을 수 없게 된다. 이러한 평가변수 설정 문제와 위암의 진행 단계(EGC, AGC)에 대한 구분이 중요하지 않다는 내시경 전문의의 임상적 의견에 따라 위암을 검출하는 것만으로 설계를 변경하였다.

우월성 검정에서 비열등성 검정으로 변경하여 위암영상검출·진단보조소프트웨어를 사용하더라도 의료진 혼자 위내시경을 분석하는 것보다 열등하지 않음을 입증하고자 하였다.

식약처 통계팀에 따르면, 비열등성 한계점을 선행연구를 토대로 설정하고자 할 경우, 통계적으로 적절한 방법은 하한값이 아닌 상한값을 설정하는 것이다. 따라서, 비열등성 검정은 초기 설계했던 대로 우월성 검정으로 다시 변경하여 위암영상검출·진단보조소프트웨어의 민감도 및 특이도를 산출한 결과를 의료진의 판독 결과와 비교하는 것이 아닌 선행연구를 통해 설정한 값과 비교하여 우월성을 입증하도록 설계하였다. 우월성 검정으로 변경하였기에, 시험군의 신뢰구간이 대조군의 신뢰구간보다 클 때 시험군이 대조군에 비해 통계적으로 유의하게 우월함을 보일 수 있다.

표 29. 임상시험 디자인 변경 과정

임상시험 디자인	1차	2차	3차
군 설정	대조군	대조군	단일군
시험군	소프트웨어의 지원을 받은 의료진	소프트웨어	소프트웨어
대조군	소프트웨어의 지원을 받지 않은 의료진	의료진	-
군 배치 방법	평행설계	평행설계	평행설계
눈가림	단일 눈가림 (평가자)	단일 눈가림 (평가자)	삼중 눈가림 (피험자, 표본 데이터 선정 및 배정자, 참조표준 구축자)
비교 검정 방법	우월성	비열등성	우월성

#### 나. 표본 데이터 선정

후향적 임상시험으로 임상시험에 참여하여 시험기기 또는 대조기기를 적용받는 피험자는 표본 데이터로 대체된다.<sup>14-16</sup> 본 연구에서는 표본 데이터 산출 과정에서 편의상 데이터 수(개)로 서술하였지만, 실제로는 위내시경 이미지가 아닌 대상자 수를 기준으로 산출하였으므로 산출된 값은 대상자(명)에 적용된다. 위내시경을 촬영한 대상자별로 한 개 이상의 이미지를 가지고 있으므로, 동일한 대상자의 다른 위내시경 이미지가 표본 데이터를 구축하는 데에 사용되었다.

연령은 만 19세 이상인 성인을 대상으로 설정하였다. 위내시경 및 조직검사 결과를 확인하여 위암의 유무를 파악할 수 있는 데이터를 선정 기준으로 정하였다. 위내시경 및 조직검사 결과 위암이 확진된 경우 양성으로 분류되고, 위암이 아니거나 양성 병변이 있는 경우 음성으로 분류된다. 선정 기준을 모두 만족하는 데이터에 한하여 표본 데이터로 선정하였다.

위암 환자의 데이터 중 병리학적 결과가 없는 경우는 위암 유무 판독이 불가하므로 제외하였다. 위절제술을 시행 받은 환자의 데이터도 제외 기준에 해당한다. 또한, 책임연구자 또는 기타 다른 연구자가 판단했을 때 임상시험에 부적절한 것으로 보이는 데이터도 제외되었다. 이러한 제외 기준에 하나라도 해당하는 경우 표본 데이터로 수집되지 않았다.



표 30. 선정 및 제외 기준

선정 기준	제외 기준
1. 위내시경 검사를 시행한 만 19세 이 상 성인	1. 병리학적 결과가 없는 위암 환자의 데이터
2. 의무기록 결과(위내시경 및 조직검사 결과지)를 통해 위암 유무 확인이 가능한 경우	2. 위절제술을 시행 받은 환자의 데이터
가. 양성: 위내시경 및 조직검사 결과 위암이 확진된 환자의 데이터	3. 책임연구자 또는 기타 연구자에 의해 임상시험 참여가 부적절하다고 판단 되는 경우
나. 음성:	
1) 위내시경 검사 결과 위암이 아 니거나	
2) 조직검사 결과 위암은 아니지 만, 양성 병변(용종(polyp), 점 막하종양(SMT), 궤양.ulcer)) 이 있는 것으로 확인된 대상자 의 데이터	

우월성 임상시험은 시험군의 결과를 내시경 전문의의 결과와 비교하여 우월함을 입증해야 한다. 시험군의 민감도, 특이도의 신뢰구간의 하한이 최소 민감도, 최소 특이도보다 클 때 우월하다고 볼 수 있다. 최소 민감도, 최소 특이도는 선행연구에 근거한 내시경 전문의의 위내시경 영상 위암 유무 판독에 대한 민감도, 특이도로 설정하였다.

위내시경 검사 경력의 평균이 6.7년인 내시경 의사가 위내시경 영상을 위암과 비위암으로 분류한 선행연구에서 각 위내시경 의사의 판독 민감도는 0.932(0.849–0.978), 0.743(0.628–0.848), 0.689(0.571–0.792), 특이도는 1.00(0.971–1.00), 0.873(0.802–0.9), 0.897(0.830–0.944)의 결과를 보였다.<sup>5</sup>

10년 이상 경력의 Senior 3명과 1~2년 경력의 Junior 5명이 조기 위암과 비

위암으로 분류한 다른 선행연구에서 민감도는 각각 0.846 (0.796–0.888), 0.669 (0.608–0.700), 0.785 (0.730–0.833), 0.650 (0.589–0.708), 0.677 (0.616–0.733), 0.654 (0.593–0.712), 0.838 (0.787–0.881), 0.638 (0.576–0.696)이고, 특이도는 0.699 (0.639–0.754), 0.803 (0.749–0.850), 0.724 (0.665–0.777), 0.816 (0.763–0.861), 0.803 (0.749–0.850), 0.816 (0.763–0.861), 0.582 (0.519–0.643), 0.824 (0.772–0.868)의 결과를 보였다.<sup>22</sup>

표 31. 선행연구 민감도 및 특이도 결과

선행연구	의료진	민감도 (95% CI)	특이도 (95% CI)
Cho, B. J. <sup>5</sup>	Endoscopist 1	0.932 (0.849–0.978)	1.000 (0.971–1.000)
	Endoscopist 2	0.743 (0.628–0.848)	0.873 (0.802–0.9)
	Endoscopist 3	0.689 (0.571–0.792)	0.897 (0.830–0.944)
Hu, H. <sup>22</sup>	Senior 1	0.846 (0.796–0.888)	0.699 (0.639–0.754)
	Senior 2	0.669 (0.608–0.700)	0.803 (0.749–0.850)
	Senior 3	0.785 (0.730–0.833)	0.724 (0.665–0.777)
	Junior 1	0.650 (0.589–0.708)	0.816 (0.763–0.861)
	Junior 2	0.677 (0.616–0.733)	0.803 (0.749–0.850)
	Junior 3	0.654 (0.593–0.712)	0.816 (0.763–0.861)
	Junior 4	0.838 (0.787–0.881)	0.582 (0.519–0.643)
	Junior 5	0.638 (0.576–0.696)	0.824 (0.772–0.868)

두 연구를 종합해 보았을 때 경력에 따라 편차가 발생하고, 실험 환경 또는 데이터 수가 동일하지 않기 때문에 차이가 발생하지만, 의료진의 민감도와 특이도의 95% 정확신뢰구간의 상한의 평균은 0.8 이상으로 추정할 수 있다. 또한, FDA 자료에서 CT 영상이나 Chest CT, Chest X-ray 영상에서 이상 병변을 검출하는 분류 모델 기반의 인공지능 소프트웨어를 대상으로 최소 목표 수치를 0.8로 설정하고 있는 것을 확인하였다.<sup>23-25</sup> 따라서, 임상시험의 최소 민감도와 최소 특이도를 합당하다고 판단되는 수치인 0.8로 설정하였다. 95% 신뢰구간은 Clopper-Pearson 방법을 통해 정확신뢰구간을 산출하여 신뢰구간의 상한과 하한 값을 구하였다(표 32).

표 32. Clopper-Pearson의 정확신뢰구간

Lower	Upper
$B\left(\frac{\alpha}{2}; x, n - x + 1\right)$	$B\left(1 - \frac{\alpha}{2}; x + 1, n - x\right)$

$\alpha$ : 0.05

x: 맞은 데이터 수

n: 테스트 데이터 수

B: Beta 분포

표본 데이터 수를 산출하기 위해 최소 민감도, 최소 특이도를 정해야 할 뿐만 아니라 예상되는 시험군의 결과인 추정 민감도, 추정 특이도를 설정해야 한다. 제조사 내부 성능 평가 결과의 95% 정확신뢰구간을 고려하여 시험군에 대해 추정 민감도를 0.856, 추정 특이도를 0.844로 설정하였다. 내부 성능 평가는 인공지능의 학습에 사용되지 않은 테스트 데이터로 수행되었다.



표 33. 내부 성능시험 결과

민감도 (95% CI)	특이도 (95% CI)
0.820 (0.779–0.856)	0.880 (0.844–0.910)

위내시경 영상 분석 인공지능 소프트웨어의 민감도 우월성 검정을 위해 귀무가설( $H_0$ )은 최소 민감도와 추정 민감도가 같다고 설정하였다(표 34). 이 가설에 근거하여 PASS (version 15, NCSS, LLC. Kaysville, Utah, USA) 프로그램을 이용하여 양성 표본 데이터 수를 산출하였다.

표 34. 민감도 우월성 검정 가설

귀무가설( $H_0$ )	대립가설( $H_1$ )
$P_0 = P_1$	$P_0 \neq P_1$

$P_0$ : 최소 민감도

$P_1$ : 추정 민감도

$H_0$ : 최소 민감도와 추정 민감도는 같다.

$H_1$ : 최소 민감도와 추정 민감도는 같지 않다.

최소 민감도, 최소 특이도가 설정되어 있는 경우 산출식에 따라 데이터 수를 산출할 수 있다(표 35).<sup>26</sup>



표 35. 데이터 수 산출식

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2 \times p_1(1-p_1)}{(p_1 - p_0)^2}$$

민감도, 특이도를 일차 유효성 평가변수로 설정하였다. Co-primary endpoint를 설정했으므로 전체 검정력이 0.8을 넘기 위해 검정력을 0.9, 유의수준을 0.05로 가정하였다. 유의수준( $\alpha$ )=0.05, 검정력( $1-\beta$ )=0.9 일 때, 위 산출식에 따라 413개의 데이터 수가 산출된다. 탈락률 20%를 고려하여 양성 표본 데이터 수는 516개로 설정하였다.

특이도 기준에서 시험군의 우월성 검정을 위한 가설은 다음과 같다(표 36).

표 36. 특이도 우월성 검정 가설

귀무가설( $H_0$ )	대립가설( $H_1$ )
$P_0 = P_1$	$P_0 \neq P_1$

$P_0$ : 최소 특이도

$P_1$ : 추정 특이도

$H_0$ : 최소 특이도와 추정 특이도는 같다.

$H_1$ : 최소 특이도와 추정 특이도는 같지 않다.

민감도와 마찬가지로 특이도 유의수준( $\alpha$ )=0.05, 검정력( $1-\beta$ )=0.9 일 때, 산출식(표 35)에 따라 음성 표본 데이터 수는 715개가 산출되고, 탈락률 20%를 고려하여 894개로 설정하였다.



최종적으로, 산출한 결과값을 토대로 양성 표본 데이터 516개, 음성 표본 데이터 894개로 총 1410개의 표본 데이터를 목표 표본 데이터 수로 설정하였다.

표 37. 표본 데이터 수

	산출 결과	탈락률 고려
양성 표본 데이터 수	413	516
음성 표본 데이터 수	715	894
총 데이터 수	-	1410

표본 데이터 수집 및 선정 시에 임의로 난이도 조절이 이루어질 수 있는 편향(Bias) 가능성을 예방하기 위해 스크리닝 담당자가 내시경 이미지를 미리 판독하지 않고 의무기록(Electronic Medical Record, EMR)만을 확인하여 선정 기준에 해당하고 제외 기준에 해당하지 않는 대상자를 선정하도록 하였다. 주관적인 판단을 배제하고 의무기록만을 확인하여 구축한 대상자 풀(Pool)에 일련의 등록번호를 부여하고, 등록번호 순서대로 위내시경 검사 및 조직검사를 통해 위선암으로 확진된 환자 516명, 위내시경 검사 또는 조직검사에서 위선암 소견이 관찰되지 않은 대상자 894명을 선별한다.

선별하는 과정에서 위암 유무와 더불어 하위그룹 정보 또한 기록하도록 하였다. 하위그룹은 내시경 전문의의 임상적 의견에 따라, 위암인 경우 조기 위암, 진행성 위암, 위암이 아닌 경우 용종(Polyp), 점막하 종양(Submucosal Tumor, SMT), 궤양(Ulcer), 병변 없음(Clean)으로 분류하였다.

표 38. 위암 및 비위암 하위그룹

위암	비위암
• 조기 위암(EGC)	• 용종(Polyp)
• 진행성 위암(AGC)	• 점막하 종양(SMT)
	• 궤양(Ulcer)
	• 병변 없음(Clean)

표본 데이터를 인공지능 소프트웨어에 적용한 후 위암 유무를 판독 결과를 기록하고, 모든 표본 데이터에 대해 판독을 수행한 후 임상시험기관에서 유효성 및 안전성에 대한 평가를 받게 된다. 인공지능에 기반한 의료 기기 학습 및 테스트 목적으로 사용된 환자 및 대상자의 데이터는 표본 데이터 구축하는 과정에서 제외되었다.

## 2. 위내시경 영상 분석 인공지능 소프트웨어 임상적 유효성 평가

### 가. 유효성 평가 분석군

유효성 평가 분석은 FAS(Full Analysis Set)군과 PP(Per Protocol)군을 기본으로 하여 수행하고, 우월성 검정 임상시험이므로 보수적인 판단을 위해 유효성 평가 변수에 대해 FAS군을 주분석군으로 하되, PP군에 대한 결과도 확인하여 유효성에 대해 종합적으로 판단하도록 하였다.

FAS 분석군은 임상시험용 의료기기를 모두 적용하고 참조표준으로 구축된 모든 표본 데이터를 대상으로 한다. 임상시험에서 무작위 배정된 모든 표본 데이터를 분석에 포함시켜야 한다는 ITT(Intention To Treat) 원칙을 가장 근접하고 완전하게 적용할 수 있다. PP 분석군은 FAS 분석군 중에서 중대한 임

상시험계획서 위반 없이 임상시험이 완료된 표본 데이터를 대상으로 한다. 선정 및 제외 기준을 위반하거나 이차 유효성 평가변수가 누락하는 경우 등이 중대한 임상시험계획서 위반에 해당된다. 제외되는 표본 데이터가 존재하므로 편향이 발생할 수 있는 문제점이 있다.<sup>20</sup>

결측치가 발생하는 경우도 고려해야 하지만, 시험군이 참조표준으로 구축한 표본 데이터의 위암 여부를 판단하는 과정에서 예측 결과가 출력되기 때문에 결측이 발생할 가능성은 없다. 만약, 시스템 오류로 인해 판독에 실패한 경우 소프트웨어를 재구동하여 판독을 수행하고, 기타 이유로 판독이 불가능한 경우에는 위양성 또는 위음성 처리하도록 하였다.

#### 나. 일차 유효성 평가변수

일차 유효성 평가변수는 위암 검출 결과의 민감도와 위암 검출 진단 결과의 특이도로 설정하였다. 시험군의 위암 검출 민감도 결과의 95% 정확신뢰구간의 하한이 0.800보다 크고, 시험군의 위암 검출 진단 특이도 결과의 95% 정확신뢰구간의 하한이 0.800보다 큰 것으로 나타난 경우 성공으로 간주한다.

##### (1) 위암 검출 결과의 민감도

위암 검출 결과에 대해 시험군의 민감도를 산출하고 이를 평가한다. 민감도는 실제 위암 양성 데이터에서 인공지능 소프트웨어가 양성으로 판단한 비율로 계산되며, 위암의 위치 정보에 해당하는 내시경 전문의가 생성한 위암 영역과 인공지능 모델에서 얻어진 영역의 IoU(Intersection over Union) 값이 5% 이상인 것에 대해서만 진양성으로 판단하고, 5% 미만일 경우 위음성으로 판단한다. 시험군에 대하여 기준 컷오프(cut-off) 값인

0.5 일 때의 위암 검출 결과의 민감도와 95% 정확신뢰구간을 계산한다. 민감도의 95% 정확신뢰구간의 하한이 최소 민감도인 0.800보다 크면 시험군의 민감도가 최소 민감도보다 크다고 할 수 있다.

민감도 계산식에서 진양성은 위암 병변이 있는 내시경 이미지를 양성으로 판단한 데이터 개수이고, 위음성은 위암 병변이 있는 내시경 이미지를 음성으로 판단한 데이터 개수이다(표 8-9). GT(Ground truth)는 내시경 전문의가 생성한 위암 영역이고, P(Predicted)는 인공지능 모델에서 얻어진 병변의 영역을 의미한다.

표 39. IoU 계산식

$$IoU = \frac{GT \cap P}{GT \cup P}$$

(P: Predicted, GT: Ground Truth)

## (2) 위암 검출 진단 결과의 특이도

위암 검출 결과에 대해 시험군의 특이도를 산출하고 이를 평가한다. 특이도는 실제 음성 데이터에서 음성으로 판단되는 비율로 계산한다. 위암 검출 결과의 민감도와 동일하게 시험군에 대하여 기준 컷오프 값인 0.5 일 때의 위암 검출 결과의 특이도와 정확 95% 신뢰구간을 계산한다. 특이도 95% 정확신뢰구간의 하한이 최소 특이도인 0.800보다 크면 시험군의 특이도가 최소 특이도보다 크다고 할 수 있다.

특이도 계산식에서 진음성은 위암 병변이 없는 이미지를 음성으로 판단한 데이터 수이고, 위양성(False Positive, FP)은 위암 병변이 없는 이미지

를 양성으로 판단한 데이터 개수이다(표 8-9). 즉, 위암이 없으므로 시험군에서 아무런 위치 표시가 나타나지 않은 경우에 대해서만 진음성으로 판단하며, 위치 정보에 해당되는 내시경 전문의가 생성한 위암 영역이 없음에도 인공지능 모델에서 병변 영역을 표시한 경우 위양성으로 판단한다.

#### 다. 이차 유효성 평가변수

이차 유효성 평가변수는 위암 검출 결과의 정확도, 위암 검출 성능에 대한 컷오프(cut-off) 유효범위 분석 및 AUC, 위암 검출 결과의 질환 하위그룹에 대한 성능(민감도, 특이도, 정확도) 분석으로 설정하였다.

##### (1) 위암 검출 진단 결과의 정확도

시험군에 대해 위암 검출 진단 결과의 정확도와 정확 95% 신뢰구간을 산출하여 정확도 결과를 확인한다. 정확도는 실제 양성을 양성으로, 실제 음성을 음성으로 판정한 능력을 의미하고, 전체 데이터에서 진양성, 진음성에 대한 비율로 계산한다(표 9). 진양성은 위암 병변이 있는 이미지를 양성으로 판단한 데이터 개수, 위양성은 위암 병변이 없는 이미지를 음성으로 판단한 데이터 개수, 진음성은 위암 병변이 있는 이미지를 음성으로 판단한 데이터 개수를 의미한다(표 9).

정확도 또한 위암 병변이 있는 표본 데이터의 경우 위암의 위치 정보에 해당되는 내시경 전문의가 생성한 위암 영역과 인공지능 모델에서 얻어진 영역의 IoU 값이 5% 이상인 것을 진양성으로 처리하고, 5% 미만일 경우 위음성으로 처리한다. 위암 병변이 없는 표본 데이터의 경우 인공지능 모

델에서 아무런 위치표시가 나타나지 않았을 때 진음성으로 처리하고, 병변 영역을 표시한 경우 위양성으로 처리한다.

표 40. 정확도 계산식

---

$$\frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

---

### (2) 위암 검출 성능에 대한 컷오프 유효범위 및 AUC

컷오프 값(0.05 간격의 0.05~0.95 까지의 구간)에 따른 위암 검출 결과의 민감도와 특이도를 구하고 ROC curve 아래부분의 넓이인 AUC 면적을 계산하여 진단 성능에 대한 컷오프의 유효범위를 분석한다.

### (3) 위암 검출 결과의 질환 하위그룹 성능

위암 검출 결과의 하위그룹은 위암인 경우 조기 위암, 진행성 위암, 위암이 아닌 경우 용종, 점막하 종양, 궤양, 병변 없음으로 나뉜다(표 38). 하위그룹 각 항목에 대해 민감도, 특이도, 정확도를 산출하고 이를 평가한다.

민감도 산출하기 위해 진양성과 위음성 값을 이용한다. 조기 위암의 경우 진양성은 조기 위암 병변이 있는 이미지를 조기 위암 양성으로 판단한 데이터 개수이고, 위음성은 조기 위암 병변이 있는 이미지를 조기 위암 음성으로 판단한 데이터 개수이다. 진행성 위암의 경우 진양성은 진행성 위암 병변이 있는 이미지를 진행성 위암 양성으로 판단한 데이터 개수이고, 위음성은 진행성 위암 병변이 있는 이미지를 진행성 위암 음성으로 판단

한 데이터 개수이다.

특이도 산출에는 진음성과 위양성 값을 이용한다. 조기 위암 및 진행성 위암의 경우 진음성은 위암 병변이 없는 이미지를 병변이 없는 것으로 판단한 데이터 개수이고, 위양성은 위암 병변이 없는 이미지를 병변이 있는 것으로 판단한 데이터 개수이다. 비위암 병변인 점막하 종양, 궤양, 용종의 경우 진음성은 비위암 병변이 없는 이미지를 비위암 병변이 없는 것으로 판단한 데이터 개수, 위양성은 비위암 병변이 없는 이미지를 비위암 병변이 있는 것으로 판단한 개수이다. 병변 없음의 경우 진음성은 병변이 없는 (clean) 이미지를 병변이 없는 것으로 판단한 데이터 개수이고, 위양성은 병변이 없는 이미지를 위암 병변이 있는 것으로 판단한 데이터 개수이다.

양성 하위 질환군인 조기 위암, 진행성 위암과 음성 하위 질환군인 비위암 병변(점막하 종양, 궤양, 용종, 병변 없음)에서의 정확도 및 신뢰구간을 제시하도록 하였다(표 38, 40).

#### (4) 위암 검출 시 IoU에 따른 민감도, 특이도, AUC

IoU 값(0.05 간격의 0.05~0.50 까지의 구간)에 따른 위암 검출 결과의 민감도와 특이도를 구하고 ROC curve 아래부분의 넓이인 AUC 면적을 계산하여 최소 민감도, 최소 특이도를 만족하는 최대 IoU값을 계산한다. 컷오프 값이 0.5일 때와 동일한 방법으로 컷오프 값에 따른 위암 검출 결과의 민감도와 특이도를 구하고 ROC curve 아래부분의 넓이인 AUC 면적을 계산하여 진단 성능에 대한 컷오프의 유효범위를 분석한다.

### (5) 위암 검출 결과의 발병 위치 하위그룹 성능

위암이 발견된 위 내 세부 위치는 상부 1/3, 중부 1/3, 하부 1/3이며, 상부 1/3은 위저부, 분문부, 상부체부를 포함하고, 중부 1/3은 중부체부, 하부체부, 위각부를 포함하며, 하부 1/3은 전정부, 유문부를 포함하는 것으로 하였다(표 41).

위암이 발견된 위치별 하위그룹인 상부 1/3, 중부 1/3, 하부 1/3에 대해 민감도, 특이도, 정확도를 산출하고 이를 평가한다(표 8-9, 40-41).

표 41. 병변의 위치별 해당 부위

상부 1/3	중부 1/3	하부 1/3
• 위저부	• 중부체부	• 전정부
• 분문부	• 하부체부	• 유문부
• 상부체부	• 위각부	



### 3. 위내시경 영상 분석 인공지능 소프트웨어 임상적 안전성 평가

인공지능 소프트웨어 안전성 평가 변수로는 이상사례를 모니터링하는 것으로 설정하였다. 본 연구에서 설계한 임상시험은 소프트웨어를 사용하는 후향적 임상시험이므로 제품이 사람에게 직접적으로 가해지지 않는다. 따라서 피험자에게 나타나는 부작용 및 안전성의 위해요소는 없다고 볼 수 있다.

이상사례(Adverse Event, AE)란 임상시험 중 피험자에서 발생한 모든 의도하지 않은 증후(症候, sign, 실험실 실험 결과의 이상 등을 포함한다), 증상(症狀, symptom) 또는 질병을 말하며, 해당 임상시험용 의료기기와 반드시 인과 관계를 가져야 하는 것은 아니다.<sup>27</sup> 이상사례의 위해정도는 경미(Mild), 중증(Moderate), 심각(Severe)으로 분류한다(표 42).<sup>20</sup>

표 42. 이상사례 위해정도

위해정도(Severity)	설명
경미 (Mild)	<ul style="list-style-type: none"><li>• 일시적이거나 경미한 불편함이 있는 경우</li><li>• 활동에 제약 없이 정상적인 일상생활(기능)을 저해하지 않고 최소한의 불편을 야기하며 피험자가 쉽게 견딜 수 있는 경우</li></ul>
중증 (Moderate)	<ul style="list-style-type: none"><li>• 중간 정도의 활동상 제약이 있고, 약간의 도움이 필요 할 수도 있는 경우</li><li>• 의학적 중재나 치료가 필요하지 않을 수도 있고 최소한으로 필요할 수도 있는 경우</li></ul>
심각 (Severe)	<ul style="list-style-type: none"><li>• 활동 상 제약이 있고, 항상 도움을 필요로 하는 경우</li><li>• 의학적 중재/치료, 입원 가능성 있는 경우</li></ul>

## IV. 고찰

위암은 국내에서 흔한 암으로 국가 차원에서 국가암검진사업의 일환으로 위암 검진을 시행하고 있다. 검진 방법 중 내시경 검사를 꼽을 수 있다. 위내시경 검사는 위점막의 형태학적 변화를 발견할 수 있고, 양성 및 악성을 감별하며, 종양의 형태 또는 침습 정도에 대해 식별을 도와줌으로써 위암 진단에 있어 그 가치가 크게 인정되고 있으며 조기 위암 발견에 상당이 유용하다.<sup>9,10</sup> 위내시경 검사 수요가 점차 늘어날 것으로 보이고, 위내시경 검사를 많이 시행 할수록 내시경 전문의들에게 부담이 되고 높은 정확도가 요구될 것이다.<sup>11</sup> 또한, 위내시경 연구가 발달함으로써 내시경 이미지의 수가 증가하고, 고화질화됨에 따라 육안으로 관찰해야 하는 의료진의 진찰 시간과 육체적 피로도가 증가하였다.<sup>12</sup> 위암을 검출하고 의료진의 진단을 보조해주어 위내시경 검사를 수행하는 의료진의 부담을 덜어줄 수 있도록 위내시경 영상 분석 인공지능 소프트웨어의 안전성 및 유효성 평가를 위한 프로토콜을 개발하였다.

위암영상검출·진단보조소프트웨어 품목에 대하여 국내에서 처음으로 설계한 임상시험으로 어려움이 있었다. 관련한 임상시험 사례는 많지 않았으며, 해당 품목에 대한 가이드라인이 없어 유사한 품목의 가이드라인을 참고하였다. 조사한 임상시험 사례들을 보면 시험군을 소프트웨어의 보조를 받은 내시경 전문의로 설정하여 소프트웨어의 보조를 받지 않은 내시경 전문의의 결과와 비교하고 있다. 반면, 본 프로토콜에서는 소프트웨어를 시험군으로 설정하고 참조표준과 비교하였기 때문에 이전 사례들보다 임상시험 수행이 용이하였다. 또한, 참조표준 구축 과정에서 촬영된 내시경 영상이 부족하여 위의 전체적인 구조를 관찰할 수 없는 환자의 데이터, 혼들림, 인공물, 음영 등으로 인해 위내시경 영상의 품질 또는 해상도가 너무 낮은 환자의 데이터, 파일 손상에 의

해 편독 불가 상태의 데이터가 발생한 경우 이는 탈락으로 간주하고 병변의 유무가 명확히 나타나 있는 이미지만을 사용하였기에 소프트웨어의 평가 결과가 성공 기준을 만족하는 데에 큰 문제가 없었다. 진단이 명확한 병변 외에도 다양한 이미지를 참조표준으로 구축하여 소프트웨어의 유효성을 평가할 필요가 있어 보인다. 표본 데이터 수 산출 시 내시경 이미지 수가 아닌 환자 수를 기준으로 하게 되면 환자당 이미지 수가 많이 때문에 총 이미지 수는 늘어나게 된다. 이에 대하여서도 명확한 기준이 필요할 것으로 보인다. 또한, 인공지능 소프트웨어의 유효성 평가에 있어 중요한 기준값인 최소 민감도, 최소 특이도 설정에 많은 어려움이 있을 것으로 예상된다.

위내시경 영상 분석 인공지능 소프트웨어에 대한 프로토콜 개발 과정에서 임상시험 설계는 식약처 의견에 따라 수차례 변경되었다. 국내에 위암영상검출·진단보조소프트웨어에 대한 임상시험 가이드라인이 없는 상황에서 해당 품목에 대해 임상시험 설계 시 많은 어려움이 있었고, 향후 임상시험을 설계 시에도 어려움을 겪을 것으로 보인다. 본 프로토콜 개발 과정에서 앞서 언급한 문제점 및 고찰과 관련한 내용이 담긴 가이드라인이 개발된다면 위암영상검출·진단보조소프트웨어 임상시험 설계 시 많은 도움이 될 것으로 예상된다.

내시경 검사 중에 대장 용종이나 위암을 자동으로 감지하고 해당 영역을 표시해 주는 내시경 모델은 이미 유럽, 일본 및 기타 국가에서 허가를 받았으며, 현재 많은 시스템이 개발 중이다.<sup>28</sup> 위암영상검출·진단보조소프트웨어는 의료진과 환자 모두의 삶의 질을 향상시키는 긍정적인 효과를 가져올 것이고, 본 프로토콜 개발을 통해 추후 국내에서도 위암영상검출·진단보조소프트웨어 또는 다른 관련 영상 분석 소프트웨어를 대상으로 임상시험 프로토콜을 개발하는 데에 활발히 기여할 수 있기를 기대한다. 또한, 인공지능에 기반한 영상검출 및 진단보조 소프트웨어가 활발히 연구개발되기를 기대한다.



## V. 결론

본 연구에서 개발한 프로토콜은 위내시경 검사에서 위암영상검출·진단보조 소프트웨어의 유효성을 확인하기 위해 인공지능 소프트웨어의 위암 검출 결과의 민감도와 특이도를 평가하여 우월함을 증명하고자 하였다. 위암에 대한 임상적인 배경과 다양한 분류 방법에 따른 유형을 설명하였다. 의료기기 소프트웨어 관련 해외 규격을 조사하였고, 국내에서 발간된 가이드라인을 조사하여 제시하고 있는 성능 평가 항목을 확인하였다.

임상시험 설계에 참고하고자 해외 위내시경 영상 분석 소프트웨어 임상시험 사례들을 조사하고, 국내 내시경 영상 분석 소프트웨어 임상시험 사례들을 조사하였다. 임상시험에 사용되는 임상시험 대상 의료기기에 대한 개요를 제시하였다. 표본 데이터의 선정 및 제외 기준을 제시하였고, 선행연구 결과를 바탕으로 임상시험 디자인을 고려하여 표본 데이터 수를 산출하고, 그에 대한 산출 근거를 도출하였다. 임상시험 설계 변경 사항을 단계별로 제시하였다. 유효성 평가를 위한 평가변수를 제시하였다. 일차 유효성 평가변수로 위암 검출 결과의 민감도, 위암 검출 진단 결과의 특이도를 제시하였고, 이차 유효성 평가변수로 위암 검출 진단 결과의 정확도, 위암 검출 성능에 대한 컷오프 유효 범위 분석 및 AUC, 위암 검출 결과의 질환 하위그룹에 대한 성능 분석(민감도, 특이도, 정확도)을 제시하였다. 위암 및 비위암에 대한 하위그룹은 내시경 전문의의 의견에 따라 분류하였다. 이러한 평가변수를 분석하는 통계분석 방법에 대해서도 제시하였다. 평가 결과가 성공임을 만족하는 성공 기준과 이상 사례 발생에 대한 조치를 위한 안전성 평가 방법을 도출하였다.

## 참고문헌

1. Korea Central Cancer Registry, National Cancer Center. Annual report of cancer statistics in Korea in 2019. Ministry of Health and Welfare. 2021.
2. Ministry of Health and Welfare. 4th National cancer control plan. 2021.
3. National Cancer Center. Quality guidelines of gastric cancer screening. Ministry of Health and Welfare. 2nd ed. 2018.
4. Kim JH, Nam SJ, Park SC. Usefulness of Artificial intelligence in gastric neoplasms. World Journal of Gastroenterology. 2021;27(24):3543.
5. Cho BJ, Bang CS, Park SW, Yang YJ, Seo SI, Lim H, et al. Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network. Endoscopy. 2019;51:1121–1129.
6. National Cancer Information Center. Gastric cancer. Ministry of Health and Welfare. 2016.
7. International Agency for Research on Cancer. Globocan 2020. <https://gco.iarc.fr/>.
8. Japanese Gastric Cancer Association. Japanese classification of gastric carcinoma: 3rd English edition. Gastric cancer. 2011;14(2):101–112.
9. Kim HJ, Kim JJ. 상부위장관 내시경 검사의 관찰과 기록법. The Korean Journal of Gastrointestinal Endoscopy. 2007;34(1):11–17.
10. Yoon SJ, Park KN, Kee CS, Lee MH, Hahm JS, Lee OC, et al. A Clinical Study of Early Gastric Cancer. Korean Journal of Medicine. 1994;47(3):381–386.
11. Vecchi M, Nuciforo P, Romagnoli S, Confalonieri S, Pellegrini C, Serio

- G, et al. Gene expression analysis of early and advanced gastric cancers. *Oncogene*. 2007;26(29):4284–4294.
12. Nanishi K, Shoda K, Kubota T, Kosuga T, Konishi H, Shiozaki A, et al. Diagnostic accuracy of the gastric cancer T-category with respect to tumor localization. *Langenbeck's archives of surgery*. 2020;405(6):787–796.
13. 식품의약품안전평가원. 의료기기 소프트웨어 허가·심사 가이드라인(민원인 안내서). 2019.
14. 식품의약품안전평가원. 뇌 영상검출·진단보조 소프트웨어 안전성·성능 및 임상시험계획서 평가 가이드라인 (민원인 안내서). 식품의약품안전처. 2020.
15. 식품의약품안전평가원. 대장암 영상검출·진단보조 소프트웨어 안전성·성능 및 임상시험계획서 평가 가이드라인 (민원인 안내서). 식품의약품안전처. 2020.
16. 식품의약품안전평가원. 전립선암 영상검출·진단보조 소프트웨어 안전성·성능 및 임상시험계획서 평가 가이드라인 (민원인 안내서). 식품의약품안전처. 2020.
17. 식품의약품안전평가원. 인공지능(AI) 의료기기 임상시험방법 설계 가이드라인(민원인 안내서). 식품의약품안전처. 2022.
18. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing Clinical Research. 4th ed. Wolters Kluwer Health; 2015.
19. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*. 2006;8(1):19–20.
20. Abdel-aleem SM. The design and management of medical device

clinical trials. John Wiley & Sons; 2012

21. Committee for Proprietary Medicinal Products (CPMP). Points to consider on switching between superiority and non-inferiority. British journal of clinical pharmacology. 2001;52(3):223.
22. Hu H, Gong L, Dong D, Zhu L, Wang M, He J, et al. Identifying early gastric cancer under magnifying narrow-band images with deep learning: a multicenter study. Gastrointestinal Endoscopy. 2021;93:1333–1341
23. U.S. Food & Drug Administration. 510(k) Summary-Aidoc Medical, Ltd.'s BriefCase. 2019.
24. U.S. Food & Drug Administration. 510(k) SUMMARY-AVICENNA.AI's CINA CHEST. 2021.
25. U.S. Food & Drug Administration. 510(k) SUMMARY-Behold.ai Technologies Limited's red dotTM. 2020.
26. 식품의약품안전평가원. 체외진단의료기기 임상적 성능시험 가이드라인 (민원인 안내서). 식품의약품안전처. 2020.
27. 식품의약품안전처. 의료기기 임상시험 관리기준(제24조제1항 관련). 2021.
28. Attardo S, Chandrasekar VT, Spadaccini M, Maselli R, Patel HK, Desai M, et al. Artificial intelligence technologies for the detection of colorectal lesions: The future is now. World Journal of Gastroenterology. 2020;26(37):5606 - 5616.

## Abstract

### **Development of Clinical Protocol to Evaluate Safety and Efficacy of Gastrointestinal Endoscopy Images Analysis Artificial Intelligence Software**

**Hee Yeong Choi**

Department of medical device engineering and management

The Graduate School, Yonsei University

(Directed by Professor **Sung Uk Kuh, Won Seuk Jang**)

Gastric cancer is the third most frequent disease after thyroid cancer and lung cancer in 2019. Gastrointestinal endoscopy is the most accurate test for diagnosing gastric cancer, and while directly observing the inside of the stomach, it is possible to detect lesions suspected of gastric cancer and to take a biopsy. Interest in endoscopy quality management has increased as studies have shown that there is a significant difference in performance of gastrointestinal endoscopy depending on the endoscopic experience. Artificial intelligence can be applied to endoscopic image reading in order to reduce the deviation of the endoscopic examination level.

Artificial intelligence software indicates the probability(%) of predicting gastric cancer and indicates the location of the lesion, which can assist endoscopists in endoscopic examination and help them make clinical decisions. In addition, early detection is expected to help improve the prognosis of gastric cancer patients.

Based on the regulations and guidelines required for the development of the protocol for gastrointestinal endoscopy Images Analysis Artificial intelligence software, this study presents clinical design and efficacy evaluation methods for computer aided gastric cancer image detection and diagnosis software. Through this study, it is expected to help endoscopists in diagnosing gastric cancer and to be used in the development of protocols for image analysis artificial intelligence software in the future. And I hope computer aided image detection and diagnosis software based on artificial intelligence will be actively researched and developed.

---

Key Words: gastric cancer, gastrointestinal endoscopy, clinical trial, artificial intelligence, software, sensitivity, specificity