Chest X-ray Abnormalities Detection

조원 : 류소리, 최샘이

INDEX

01
주제

주제 및 주제 선정 이유
개발 환경

02
분석 목표

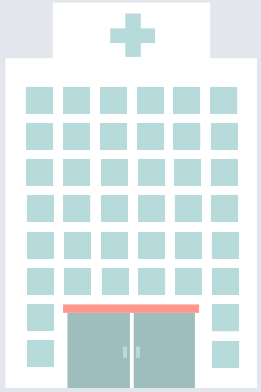
분석 목표 및
계획

03
코드

데이터 분석
시각화

04
결론

Kaggle 제출
추가 수정(공부) 방향



01-1

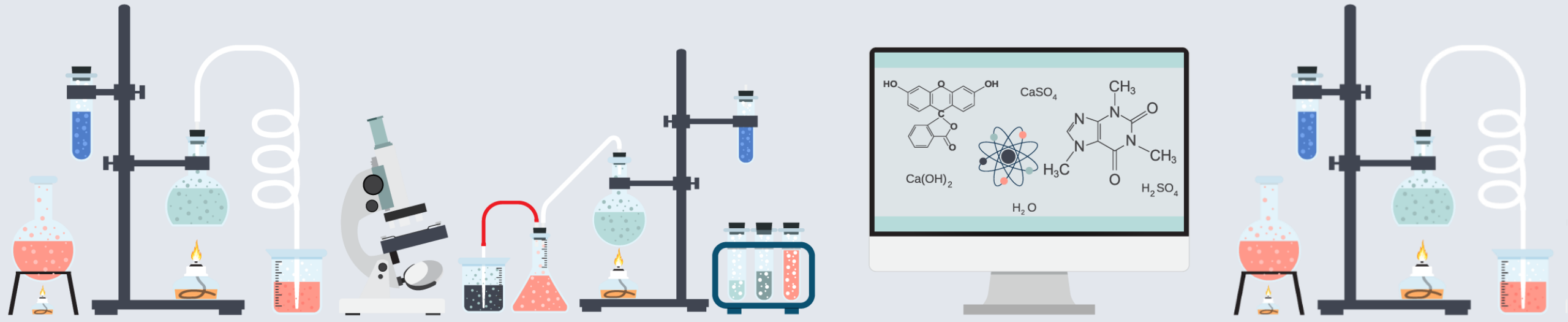
주제 및 주제 선정이유

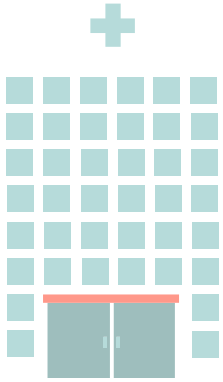
주제

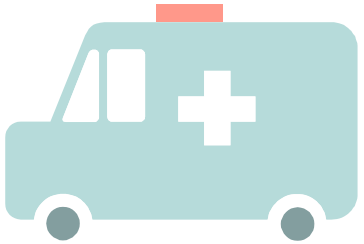
- 의사는 CT, PET scan, MRI , X-ray 와 같은 데이터를 판독하며 환자의 의학적 상태를 진단하고 치료하는데, 흉부 X-ray 의 경우 다른 부위보다 의학적 오진이 생기기 쉬운 부위이다.
- 생명과 직결된 부위인 흉부 X-ray에서 작은 size의 병변까지 보다 정확하게 식별하고 위치 파악을 할 수 있다면, 전문의사(방사선과)가 아니더라도 판독에 도움을 줄 수 있을 것이다.

선정 이유

- 이미지 데이터의 처리 및 분석에 대하여 궁금증이 있었기 때문에 프로젝트를 진행하며 알아보고 싶었습니다.
- 의료 전공 팀원이 있기에 분석 및 결과 해석이 용이 할 것 같아, 의료 이미지를 이용하는 것으로 방향을 잡았습니다.







01-2

개발환경 & 데이터 수집

01

개발환경

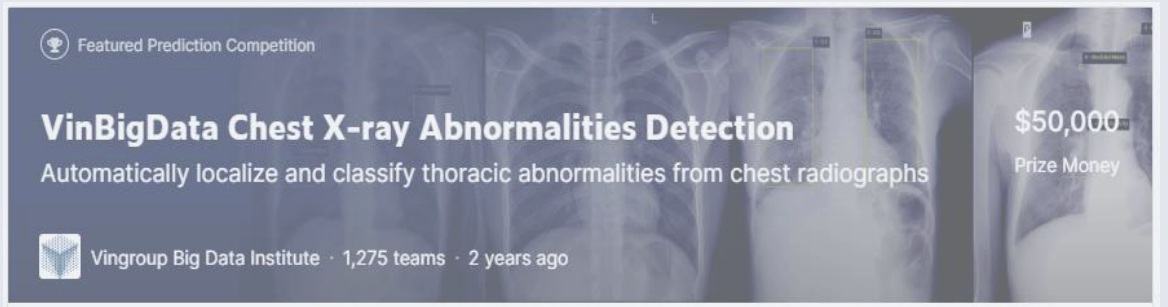
Tool	Jupyter lab Google Colab
language	Python
Library	Pandas Numpy skleran etc

02

데이터 수집

Vingroup Big Data Institute가
VinBigData의 웹 기반 플랫폼인 VinLab을 통해 수집한 이미지이며,
기초 연구를 촉진하여 새롭고 적용 가능성이 높은
기술 조사를 목표로 한 대회의 데이터 입니다.

해당 데이터는 아래 Kaggle Web Site를 통하여 데이터를 제공받았습니다.



02

분석 목표

분석 목표 및 계획

✓ 흉부의 X-ray 이미지만으로
이상 유무를 정확하게 평가할 수 있을까

✓ 흉부의 X-ray 이미지에서
14가지 유형의 병변과 한가지의 이상 없음을 파악하고
올바른 곳에 바운딩 박스를 그려보려고 한다.

03-1

Train csv 데이터 분석

Train.csv Data Analysis

[Columns]

- image_id 이미지와 매칭하기 위한 고유 식별자
- class_name 감지된 개체의 클래스 이름(혹은 No finding)
- class_id 감지된 개체의 클래스 ID
- rad_id 관찰한 방사선 전문의의 ID
- x_min bounding box의 최소 X좌표
- y_min bounding box의 최소 y좌표
- x_max bounding box의 최대 x좌표
- y_min bounding box의 최대 y좌표

	image_id	class_name	class_id	rad_id	x_min	y_min	x_max	y_max
0	50a418190bc3fb1ef1633bf9678929b3	No finding	14	R11	NaN	NaN	NaN	NaN
1	21a10246a5ec7af151081d0cd6d65dc9	No finding	14	R7	NaN	NaN	NaN	NaN
2	9a5094b2563a1ef3ff50dc5c7ff71345	Cardiomegaly	3	R10	691.0	1375.0	1653.0	1831.0
3	051132a778e61a86eb147c7c6f564dfe	Aortic enlargement	0	R10	1264.0	743.0	1611.0	1019.0
4	063319de25ce7edb9b1c6b8881290140	No finding	14	R10	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...
67909	936fd5cff1c058d39817a08f58b72cae	No finding	14	R1	NaN	NaN	NaN	NaN
67910	ca7e72954550eeb610fe22bf0244b7fa	No finding	14	R1	NaN	NaN	NaN	NaN
67911	aa17d5312a0fb4a2939436abca7f9579	No finding	14	R8	NaN	NaN	NaN	NaN
67912	4b56bc6d22b192f075f13231419dfcc8	Cardiomegaly	3	R8	771.0	979.0	1680.0	1311.0
67913	5e272e3adbdaafb07a7e84a9e62b1a4c	No finding	14	R16	NaN	NaN	NaN	NaN

67914 rows × 8 columns

Train.csv Data Analysis

Train Data Size : 67914

Train Data Columns : Index(['image_id', 'class_name', 'class_id', 'rad_id', 'x_min', 'y_min', 'x_max', 'y_max'], dtype='object')

Train Data info :

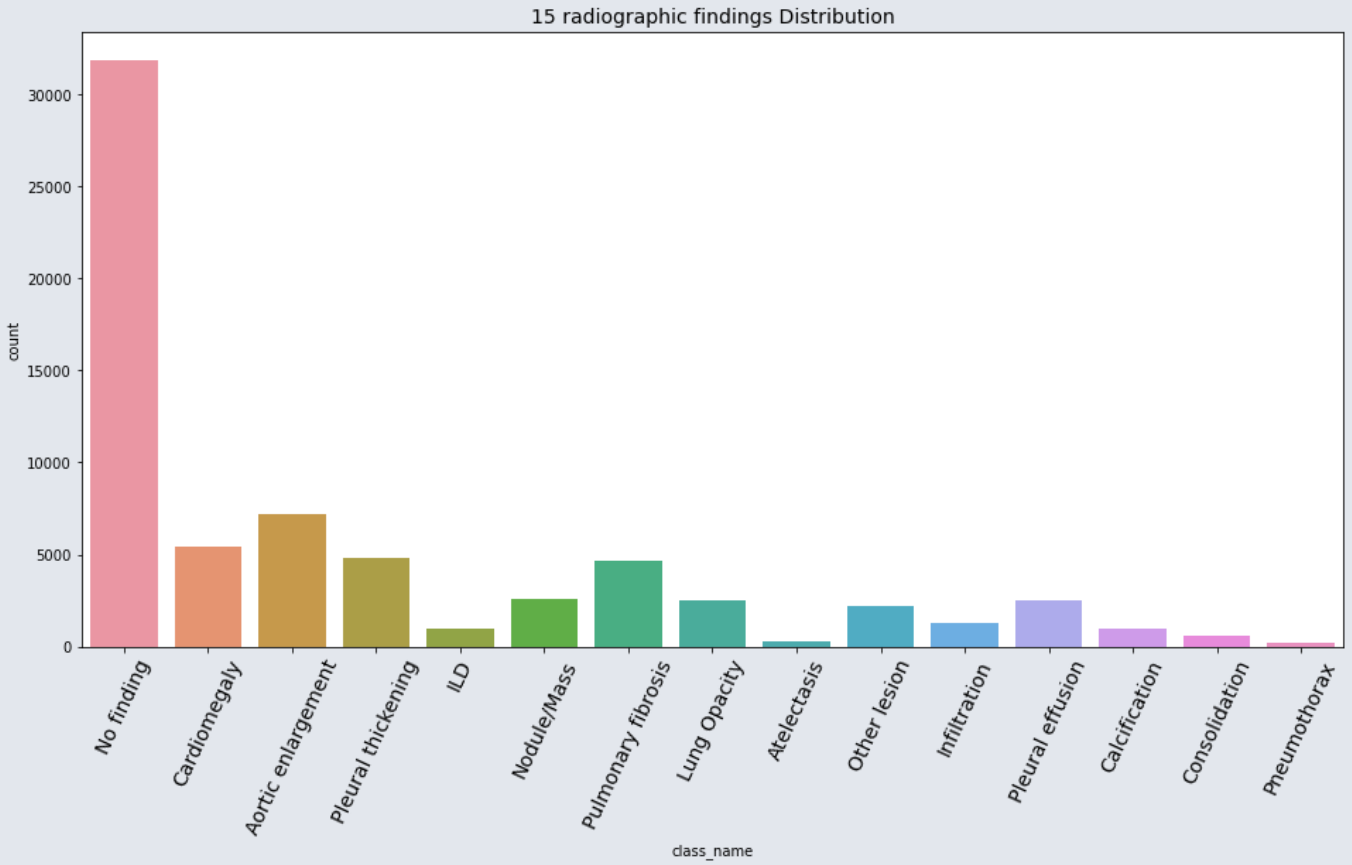
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67914 entries, 0 to 67913
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  ------      --------------  -----
0   image_id    67914 non-null  object
1   class_name  67914 non-null  object
2   class_id    67914 non-null  int64
3   rad_id      67914 non-null  object
4   x_min       36096 non-null  float64
5   y_min       36096 non-null  float64
6   x_max       36096 non-null  float64
7   y_max       36096 non-null  float64
dtypes: float64(4), int64(1), object(3)
```

총 67914개의 데이터가 담겨있으며 컬럼명은 위와 같습니다.

사이킷런에서는 문자열을 입력 값으로 처리하지 않기 때문에 Dtype의 Object형을 숫자형으로 변환해야 하는 지 살펴 봤으나
Image_id의 경우 이미지 파일과 매칭되는 string값이기 때문에 현재로서는 변환하지 않아도 될 것으로 보이며,
병명 혹은 이상 없음의 이름인 class_name의 경우 int형태의 class_id값으로 구분이 가능하기 때문에 변환하지 않으려 합니다.
방사선 전문의의 ID인 Rad_id의 경우에는 분석방향에 필요 없는 값이라 컬럼을 삭제하려 합니다.

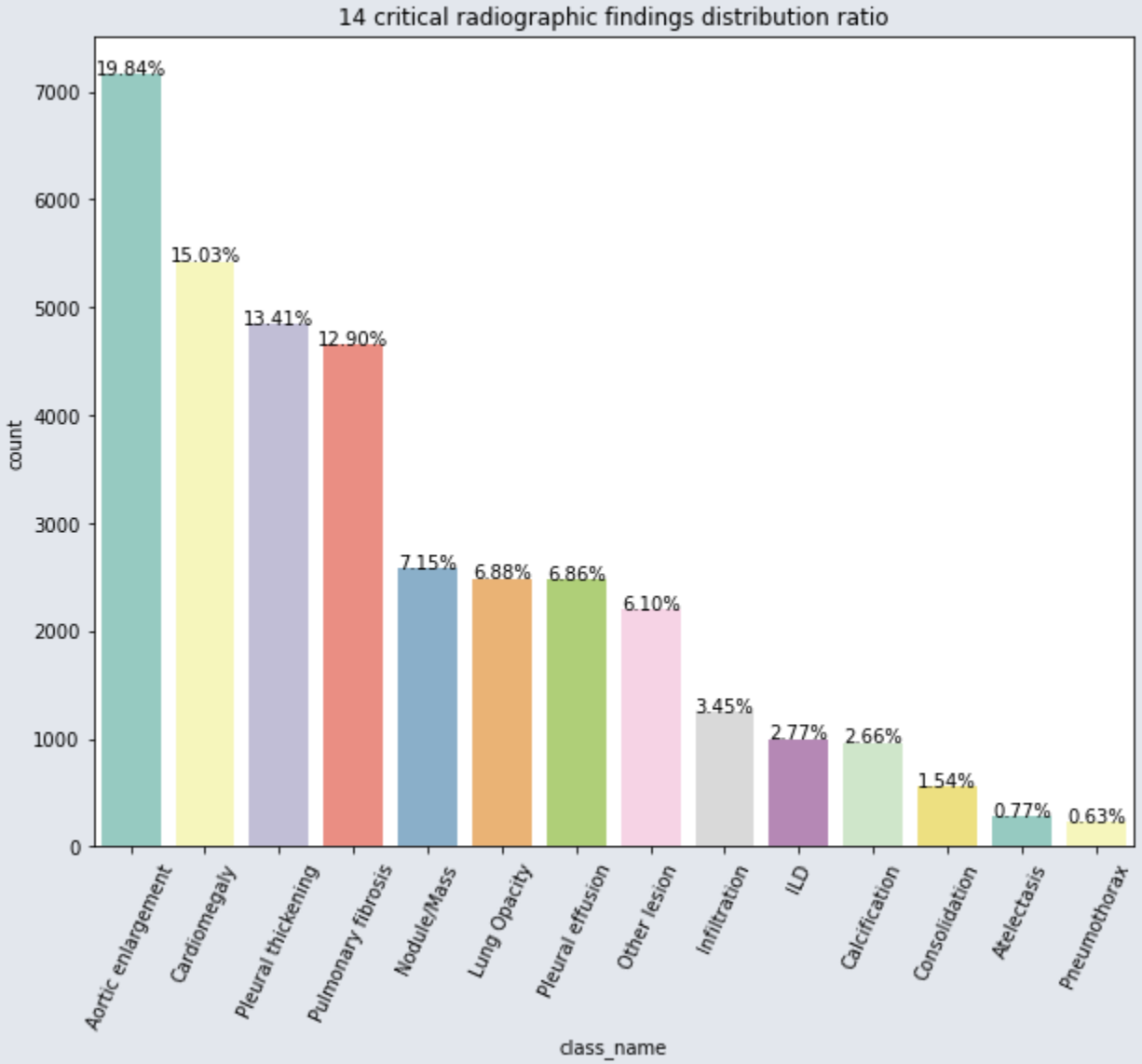
Train.csv Data Analysis

```
No finding                31818
Aortic enlargement         7162
Cardiomegaly               5427
Pleural thickening         4842
Pulmonary fibrosis         4655
Nodule/Mass                2580
Lung Opacity               2483
Pleural effusion           2476
Other lesion               2203
Infiltration               1247
ILD                        1000
Calcification               960
Consolidation               556
Atelectasis                 279
Pneumothorax                226
Name: class_name, dtype: int64
```



15가지의 병변 구분에 있어 데이터가 어떤 식으로 분포 되어있는지 확인
→ No finding 이 압도적으로 많은 것을 알 수 있다.

Train.csv Data Analysis



가장 많이 잡힌 No finding을 제외하고
14가지의 질병 조사 결과만 비율(%)과 함께 확인하였다.
→ Aortic enlargement (대동맥 확장)이 19.84%로 가장 많았고
그 다음으로는 Cardiomegaly (심장 비대) 15.03%
Pleural thickening (흉막 비후) 13.41%
순으로 많은 것을 확인하였다.

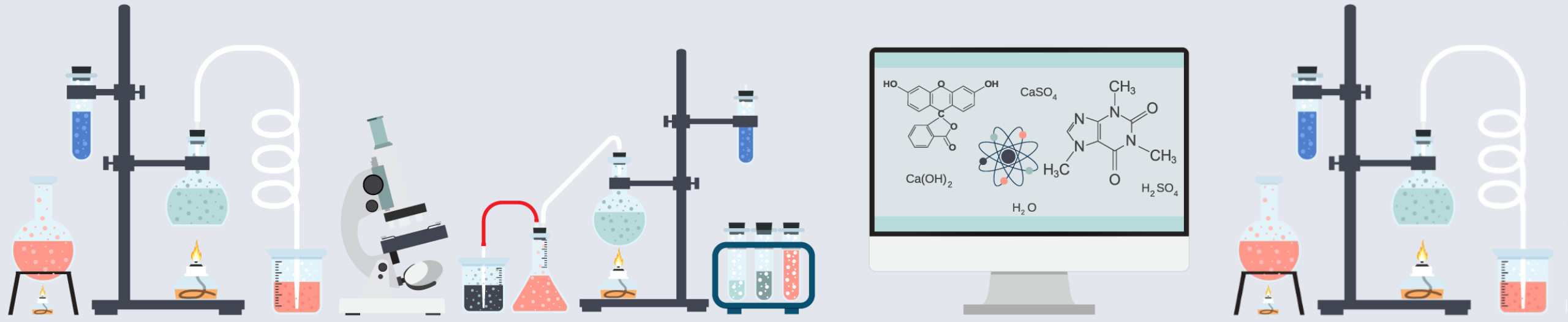
03-2

이미지 확인

먼저,

우리가 아는 일반적 이미지 확장자(.jpg .png etc)로 되어 있는 것이 아니라 DICOM이라는 확장자인 것을 확인하였다.
DICOM은 의료용 디지털 영상 및 통신 표준 확장자로 의료용 기기에서
디지털 영상표현과 통신에 사용되는 여러가지 표준을 통칭하는 확장자이다.

이 파일을 다룰 수 있도록 도와주는 Library로는 pydicom, SimpleITK 등이 있는 것으로 보이며
Pydicom의 용례가 가장 많은 것으로 보여 이를 이용하려 한다.



Sample Image

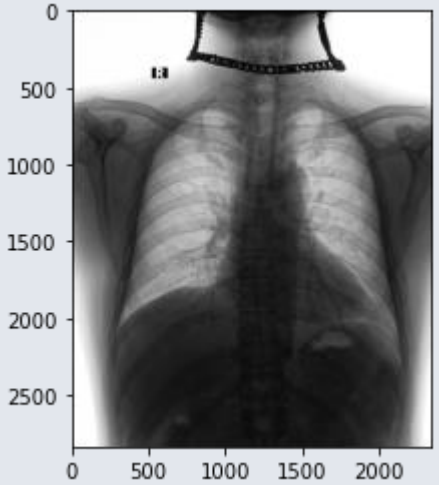


Image 출력 확인 을 위하여
Sample Image Data를 추출해 봤으며

Sample Image

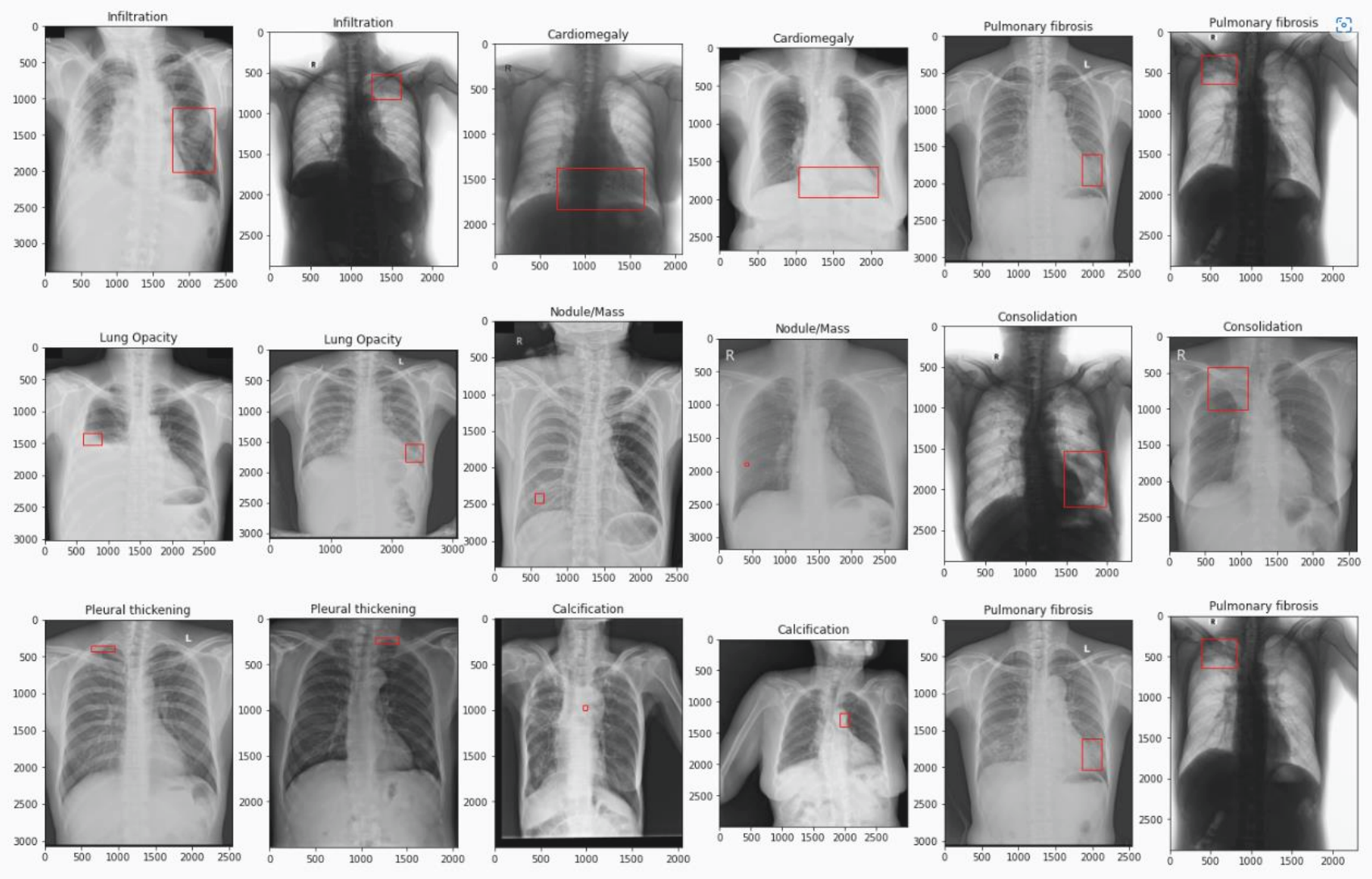


Image 출력 확인 을 위하여
Sample Image Data를 추출해 봤으며

출력이 정상적으로 되는 것으로 보여
무작위로 Class_id를 골라 2장씩 총 16장의
Sample Image를 추출한 뒤
Train Data를 활용하여
바운딩 박스를 그려주었다.
→ 올바른 정답 레이블의 모습 확인

04

추후 계획

처리

필요 없는 컬럼 삭제

이미지 데이터 변형

model

- 최대한 많은 model
- 정확도 및 측정 지표의 값을 참고하여 수정

시각화

가장 높은 정확도를 보인 모델과
가장 낮은 정확도를 보인 모델이
각각 어느 부위에 집중하여
학습했는지 확인

이 대회는 Vingroup Big Data Institute가 VinBigData의 웹 기반 플랫폼인 VinLab을 통해 수집한 이미지를 통해 기초 연구를 촉진하여 새롭고 적용 가능성이 높은 기술 조사를 목표로 한 대회이며, Kaggle Web Site를 통하여 데이터를 제공받았습니다.

INDEX

주제

주제 및 주제 선정 이유

개발 환경

학습 목표

학습 목표 및 계획

코드

코드 리뷰

결과

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

코드 분석

시각화 분석

제목을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

01
소제목을 입력하세요

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

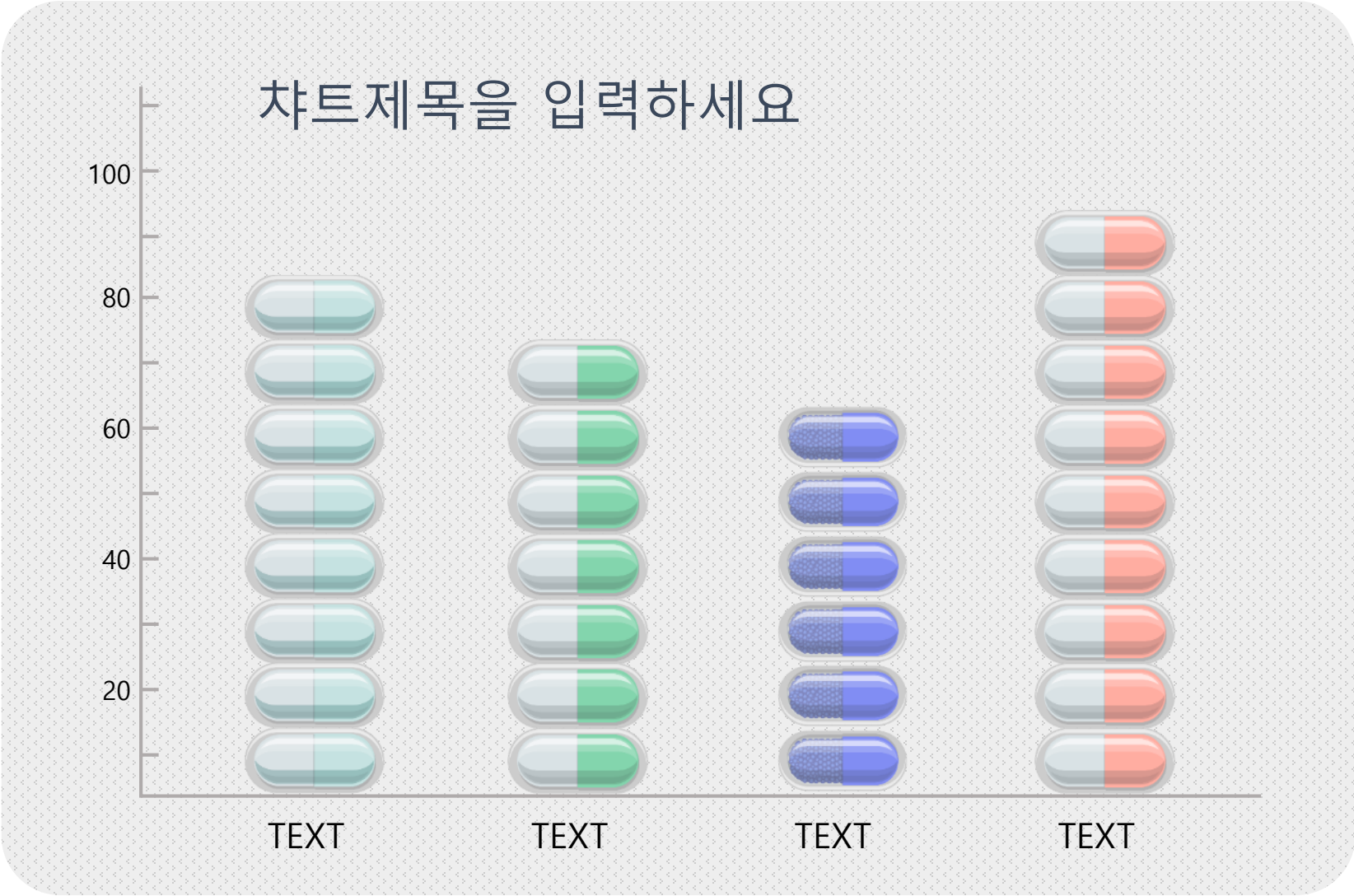
03
소제목을 입력하세요

02
소제목을 입력하세요

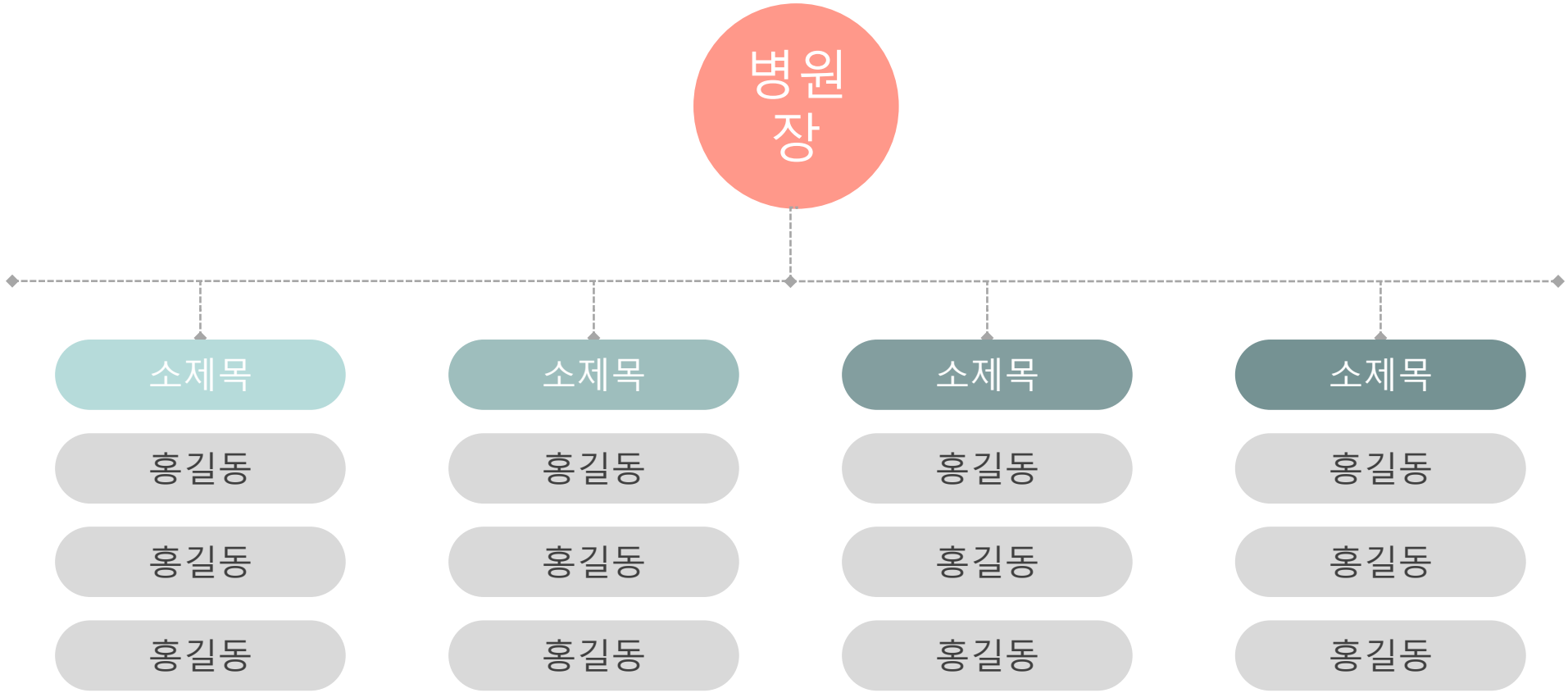
소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

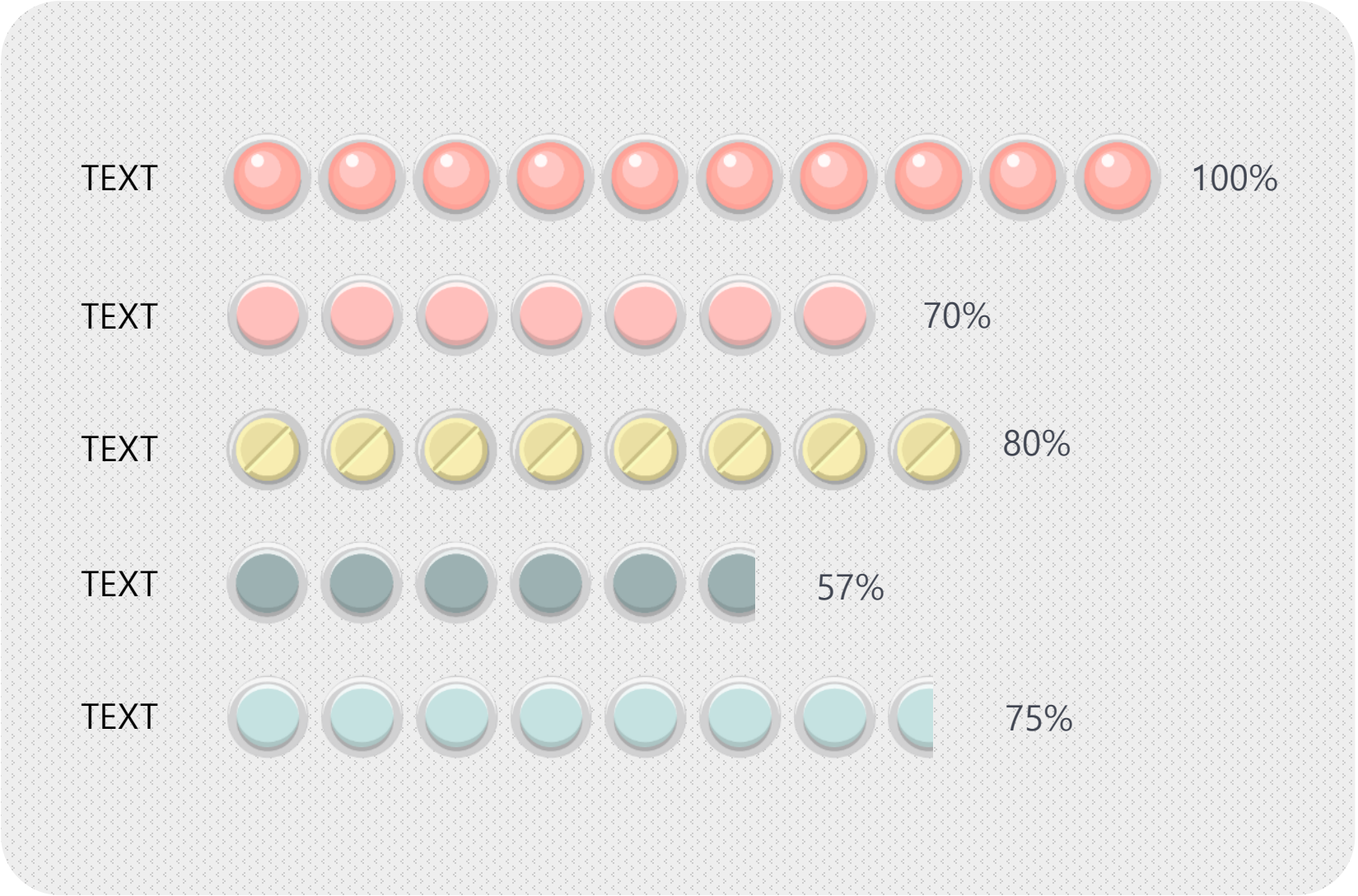
제목을 입력하세요.



제목을 입력하세요.


병원장
소제목
홍길동
홍길동
홍길동
소제목
홍길동
홍길동
홍길동
소제목
홍길동
홍길동
홍길동
소제목
홍길동
홍길동
홍길동

제목을 입력하세요.



제목을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.

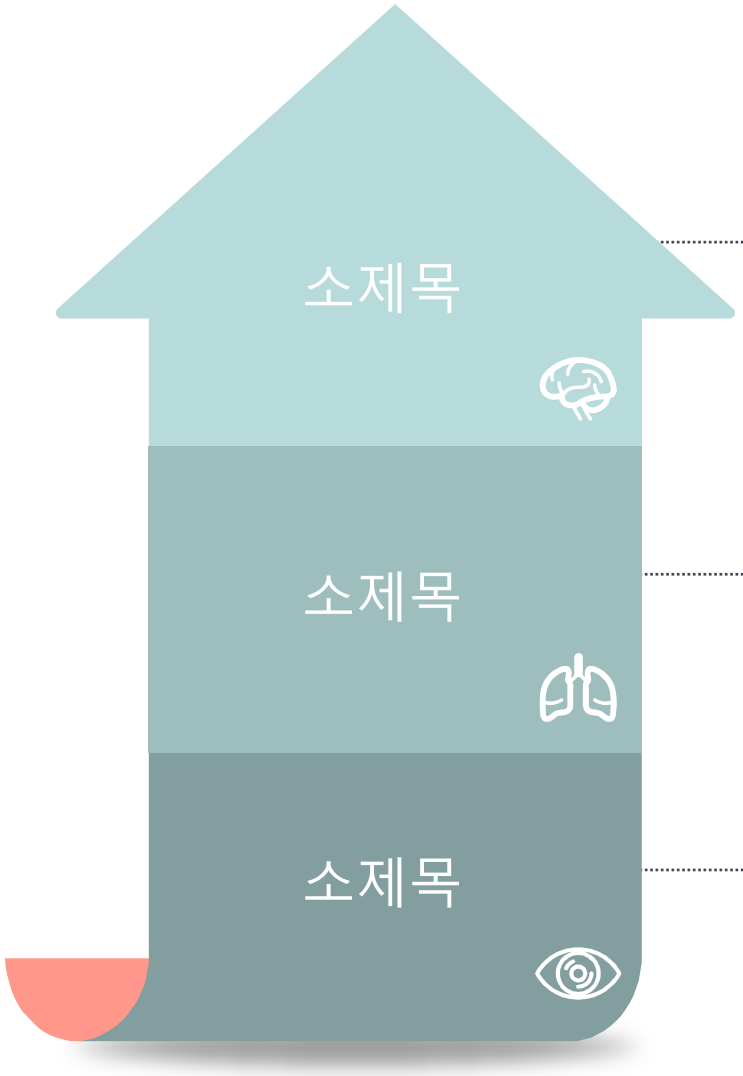
소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.

제목을 입력하세요.



01 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

02 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

03 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

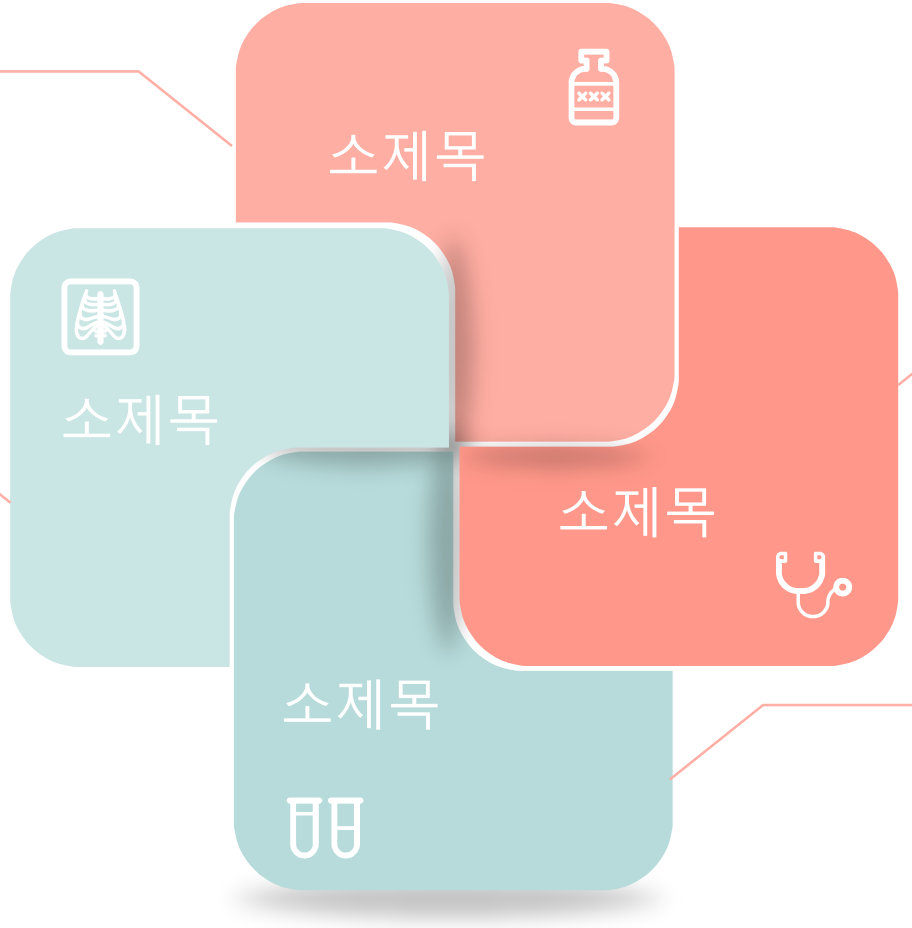
제목을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.



소제목

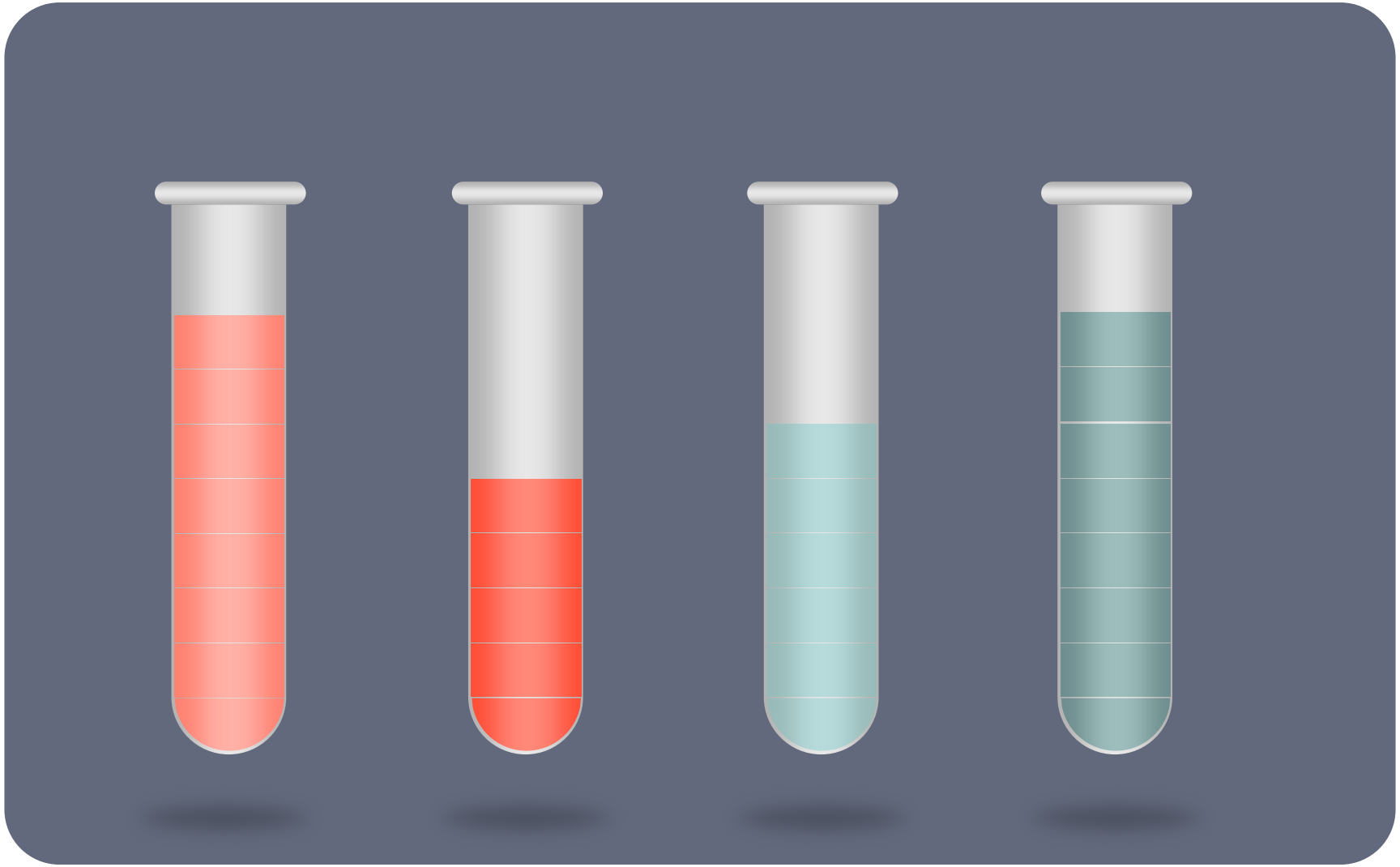
내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

Medical presentation

제목을 입력하세요.



제목을 입력하세요.

01 소제목

02 소제목

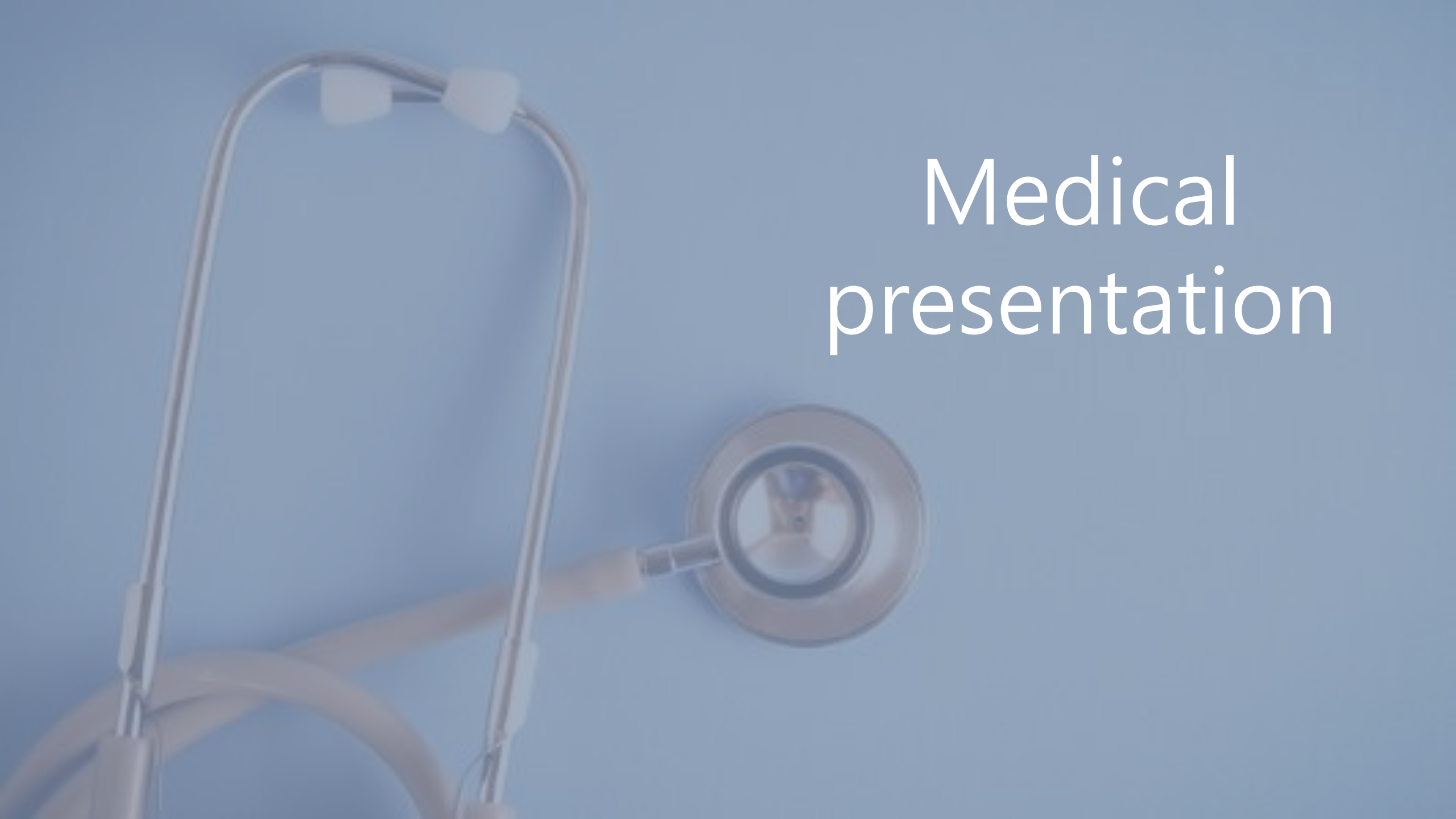
03 소제목

04 소제목

05 소제목

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.
내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.
내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.



Medical presentation

제목을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요.
내용을 입력하세요.내용을 입력하세요.

제목을 입력하세요.

01 소제목 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

소제목 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요. 02

03 소제목 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

소제목 내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요. 04

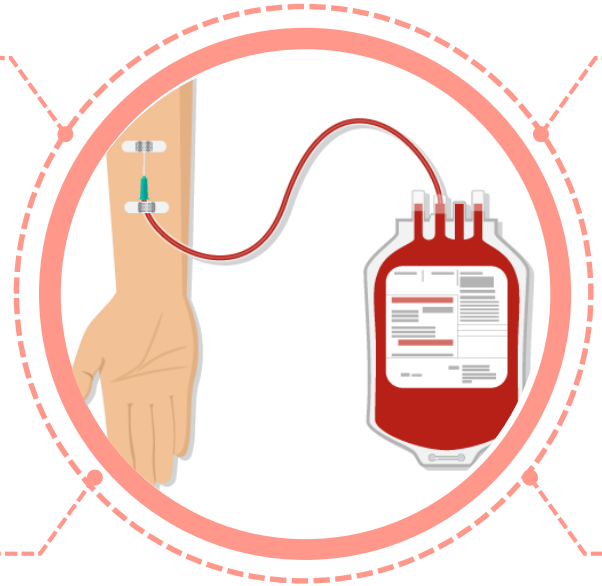
제목을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.



소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

소제목

내용을 입력하세요. 내용을 입력하세요. 내용을 입력하세요.내용을 입력하세요.

Thank you

본 문서로 할 수 있는 것

- 디자인의 변형 및 가공
- 문서의 읽기전용 파일 공유 (PDF, 이미지)
- 발표, 프리젠테이션, 제출용, 인쇄물 공유

본 문서로 하면 안되는 것

- 문서 전체 및 일부 상업적 사용 불가
- 원본 문서 파일 공유 불가
- 재판매용 배포 금지