"데이터 품질 평가기반 데이터 고도화 및
데이터셋 보정 기술 개발" 2021년도 Kick-Off 회의

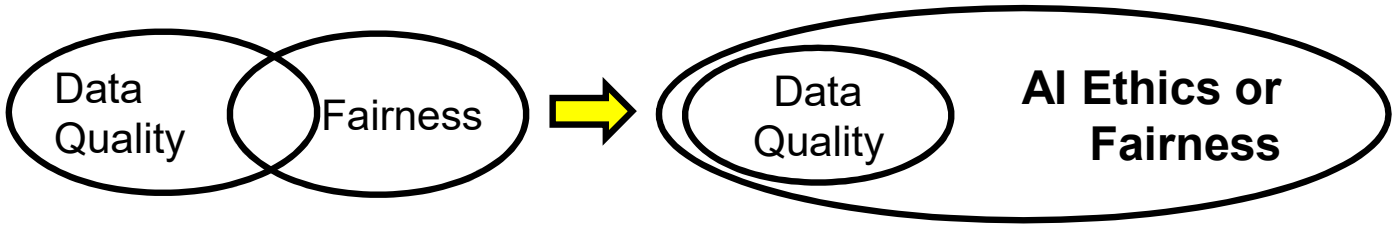
보건의료분야 진도보고

"From Algorithmic Fairness to Data Fairness in the Health Data"

QUFA
QUALITY & FAIRNESS
AEQUITAS
LUX NOSTRA

연구진행 1 – 핵심개념 정리

자료품질과 공정성에 대한 기존의 개념과 최근의 개념



윤리기준: AI의 FAST track 원칙

- 기존에는 Data quality와 AI 공정성을 별도의 개념으로 간주
- 최근 공정성이 중요한 화두가 되면서, 자료 품질은 공정성의 필요조건으로 또, 자료를 통해 산출되는 AI (algorithm)의 품질과 성능과 윤리성을 평가하는 기준으로 간주하려는 흐름 (예: AI의 FAST track원칙)
- 이러한 윤리성, 공정성이 상위범주이고, 그 안에 data quality가 보장되어야하는 개념

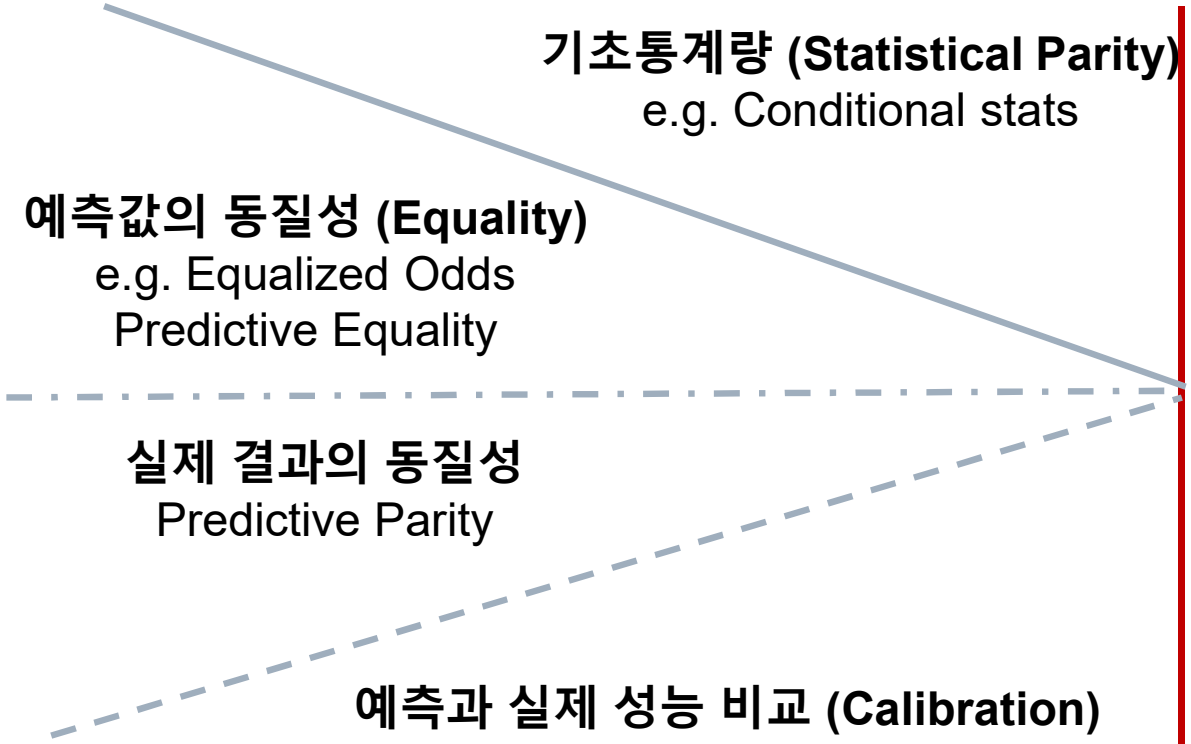
연구진행의 핵심 question과 진행방안

- **보건의료분야에서 Data 자체의 보정없이 현재의 fairness mitigation방법 적용으로 fairness특히, real world fairness를 반영하는 causal fairness index가 개선될 수 있을것인가?**
  - 보건의료 분야는 fairness와 관련된 “sensitive variables”가 이미 잘 알려져 있음
    - *Sex, Social Class, Race(Ethnicity)*
  - Sex, social determinants등은 매우 광범위한 건강결정요인 및 건강결과와 correlation을 가짐
- **실증적인 평가를 위한 data set의 구성과 평가**

보건의료 자료와 Fairness의 평가방법

- **보건의료 데이터와 분야에 적합한 Fairness평가 방법은?**
  - Fairness에 대한 논의는 초보적이며 특히 metrics가 실증적으로 평가되지 못함
  - 특히 보건의료 부문에서는 다양한 fairness metrics중에서 “인과적” metrics로 인정받고 있는 **counterfactual fairness index**값이 적절
  - 임상에 적용되고 있는 예측알고리듬의 실증적인 평가 (**calibration**)
- **문제제기:알고리듬차원의 완화가 실제 real world의 불공정성을 해결할 수 있는가?**

Fairness관련 동향 – Fairness Metrics



기초통계량 (Statistical Parity)
e.g. Conditional stats

예측값의 동질성 (Equality)
e.g. Equalized Odds
Predictive Equality

실제 결과의 동질성
Predictive Parity

예측과 실제 성능 비교 (Calibration)
e.g. Accuracy
Treatment Equality

인과적 지표 (Causal Discrimination)
Fairness Through Awareness
Counterfactual Fairness

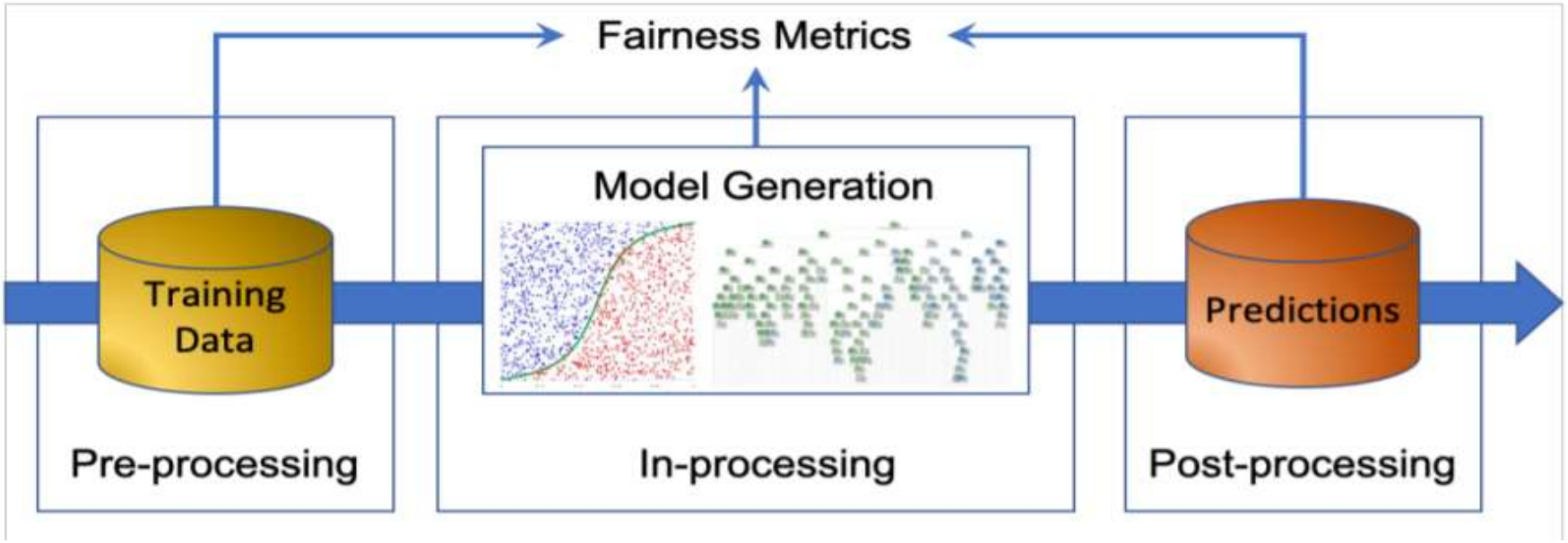
$$P(\hat{Y}_{A<-\text{남자}}(U) = y \mid X = x, A = a)$$
$$= P(\hat{Y}_{A<-\text{여자}}(U) = y \mid X = x, A = a)$$

예를 들어, 남녀간의 평가에서 다른 모든 요인이 동일할 때 모든 남자를 -> 해당되는 각각의 여자로 바꾼 결과가 동일한가

Fairness의 평가와 mitigation pipeline tools

Project	주요 Features	Source code
AIF360 (IBM)	Set of tools that provides several pre-, in-, and post-processing approaches for binary classification as well as several preimplemented datasets that are commonly used in Fairness research	https://aif360.mybluemix.net/
Fairlean (Microsoft)	Implements several parity-based fairness measures and algorithms for binary classification and regression as well as a dashboard to visualize disparity in accuracy and parity.	https://fairlearn.org/
Aequitas (open)	Open source bias audit toolkit. Focuses on standard ML metrics and their evaluation for different subgroups of a protective attribute.	http://www.datasciencepublicpolicy.org/projects/aequitas/
Responsibly	Provides datasets, metrics, and algorithms to measure and mitigate bias in classification as well as NLP (bias in word embeddings).	
Fairness	Tool that provides commonly used fairness metrics (e.g., statistical parity, equalized odds) for R projects.	
FairTest	Generic framework that provides measures and statistical tests to detect unwanted associations between the output of an algorithm and a sensitive attribute.	
Fairness Measures	Project that considers quantitative definitions of discrimination in classification and ranking scenarios. Provides datasets, measures, and algorithms (for ranking) that investigate fairness.	
Audit AI	Implements various statistical significance tests to detect discrimination between groups and bias from standard machine learning procedures.	
Dataset Nutrition Label	Generates qualitative and quantitative measures and descriptions of dataset health to assess the quality of a dataset used for training and building ML models.	
ML Fairness Gym (Google)	Part of Google's Open AI project, a simulation toolkit to study long-run impacts of ML decisions. Analyzes how algorithms that take fairness into consideration change the underlying data (previous classifications) over time	https://github.com/google/ml-fairness-gym

현재의 Bias Mitigation approach 요약 (metrics의 관점)



Reweighing
Resampling
Transformation
  Relabelling & Perturbation
Variable Blinding

Regularization
Constraint Optimization
Privileged learning
Meta Fair Classifier
Adversarial learning

Equalized Odds
Calibrated Equalized Odds
Reject Option Classification

Fairness중심의 평가를 위한 보건의료 분야 자료

- **건강보험자료 중 표본 150만명의 자료 (학술성과 및 특허제출)**
  - 건진, 문진, 의료이용 (상병내역), 보험료등급, 사망여부 등
  - 내부망 (공단직영)에서만 분석가능
  - 작년 11월 신청했으나, 아직 자료사용 대기 중 -> 타연구의 자료 접근시작
- **NIA 데이터 센터의 자료**
  - 질평가 등을 위해 반드시 사용해야하는 자료이나, 연구용으로는 부족 (질병발생결과나 사망발생등의 자료부족)
  - 품질평가 준비가 완료되는대로 데이터센터 및 플랫폼 공식적인 품질평가 추진 필요
- **기타 자료원 (건강보험자료 보완용)**
  - 질병관리청 유전체센터의 20만명 코호트자료 (행태, 대사질환등의 자료 풍부)
  - 현재 확보 중임

건강보험자료 DB 내역

구분			파일명
자격 및 보험료테이블			NSC2_BNC
출생 및 사망테이블			NSC2_BND
진료테이블	의과, 보건기관(M)	일반내역(T20)	NSC2_M20
		진료내역(T30)	NSC2_M30
		상병내역(T40)	NSC2_M40
		처방내역(T60)	NSC2_M60
	치과(D)	일반내역(T20)	NSC2_D20
		진료내역(T30)	NSC2_D30
		상병내역(T40)	NSC2_D40
		처방내역(T60)	NSC2_D60
	한방(K)	일반내역(T20)	NSC2_K20
		진료내역(T30)	NSC2_K30
		상병내역(T40)	NSC2_K40
건강검진테이블	일반건강검진(1차)	2002~2008년	NSC2_G1E_0208
		2009~2015년	NSC2_G1E_0915
요양기관테이블			NSC2_INST

RN_INDI	개인고유번호	개인고유번호(7자리), 연계코드
BTH_YYYY	출생년도	표본 대상자의 출생년도(연령: 기준년도-출생년도도 산출)
DTH_YYYYMM	사망연월	사망자의 사망월 … 통계청 사망원인 연계
COD1	사망원인1	한국표준질병·사인분류(KCD) 코드 사용
COD2	사망원인2	사망원인이 S00-T98인 경우 상세 원인 기재(V01-Y98)

영양섭취행태	Q_NTR_PRF	숫자	8	1: 채식을 주로 먹는다. 2: 채식, 육식을 골고루 먹는 편이다. 3: 육식을 주로 먹는다.
음주습관	Q_DRK_FRQ_V0108	숫자	8	1: (거의)마시지 않는다 2: 월2~3회정도 마신다 3: 일주일에 1~2회 마신다 4: 일주일에 3~4회 마신다 5: 거의 매일 마신다
1회 음주량	Q_DRK_AMT_V0108	숫자	8	1: 소주 반 병 이하 2: 소주 한 병 3: 소주 1병 반 4: 소주 2병 이상
흡연상태	Q_SMK_YN	숫자	8	1: 피우지 않는다 2: 과거에 피웠으나 지금은 끊었다 3: 현재도 피운다
(현재)하루흡연량	Q_SMK_NOW_AMT_V0108	숫자	8	1: 반갑미만 2: 반갑이상~한갑미만 3: 한갑이상~두갑미만 4: 두갑이상
(과거,현재)흡연기간	Q_SMK_DRT	숫자	8	1: 5년 미만 2: 5~9년 3: 10~19년 4: 20~29년 5: 30년 이상
흡연시작연도	Q_SMK_STRT_YR	숫자	8	YYYY * 2005년부터 적용
금연시작연도	Q_SMK_STOP_YR	숫자	8	YYYY * 2006년부터 적용
1주 운동횟수	Q_PA_FRQ	숫자	8	1: 안한다 2: 1~2회 3: 3~4회 4: 5~6회 5: 거의 매일

자료 QC관련 진행사항

- 일반적인 rule-based QC지표 정리
- 새로운 anomaly detection방법 (논의 수준)

“Deep” Anomaly Detection

- **Processing 및 분석 이전의 Data 단계에서 평가방법 필요**
  - 특히 unsupervised anomaly detection으로 labeling dependency 해결
  - 일부자료 특히 fairness key변수는 supervised anomaly detection병행
- **Deep Anomaly Detection**
  - Autoencoder(AE) and VAE
  - Adversarial Network-based anomaly detection (“**AnoGAN”, “**EGBAD”, “GANomaly”
    - *보건의료 분야는 “anomaly” (예 : 환자자료)가 희소하고 labeling이 어려운 경우가 많음*
    - *“정상”에서 train된 discriminator가 “환자”의 자료를 “fake”로 평가하도록 (fake=anomaly)*