The ATEN Framework for creating the *Realistic Synthetic Electronic Health Record*

Scott McLachlan

Kudakwashe Dube, Thomas Gallagher, Bridget Daley,
Jason Walonoski

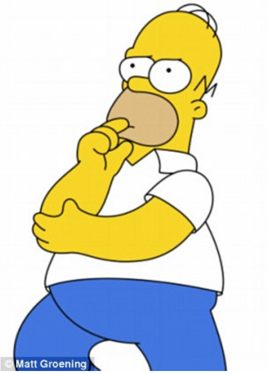
Massey University

Queen Mary
University of London

Introduction

A Cautionary Tale in which:

- A Health IT provider poorly compensates for inadequacy
- Some midwives unknowingly go where they weren't supposed to
- An Information Science graduate student gets an idea…



...and spends a year contemplating pseudorealism!

Synthetic Data Generation

Generating synthetic data seems easy…

Generating good synthetic data is far more difficult

Synthetic Data Generation

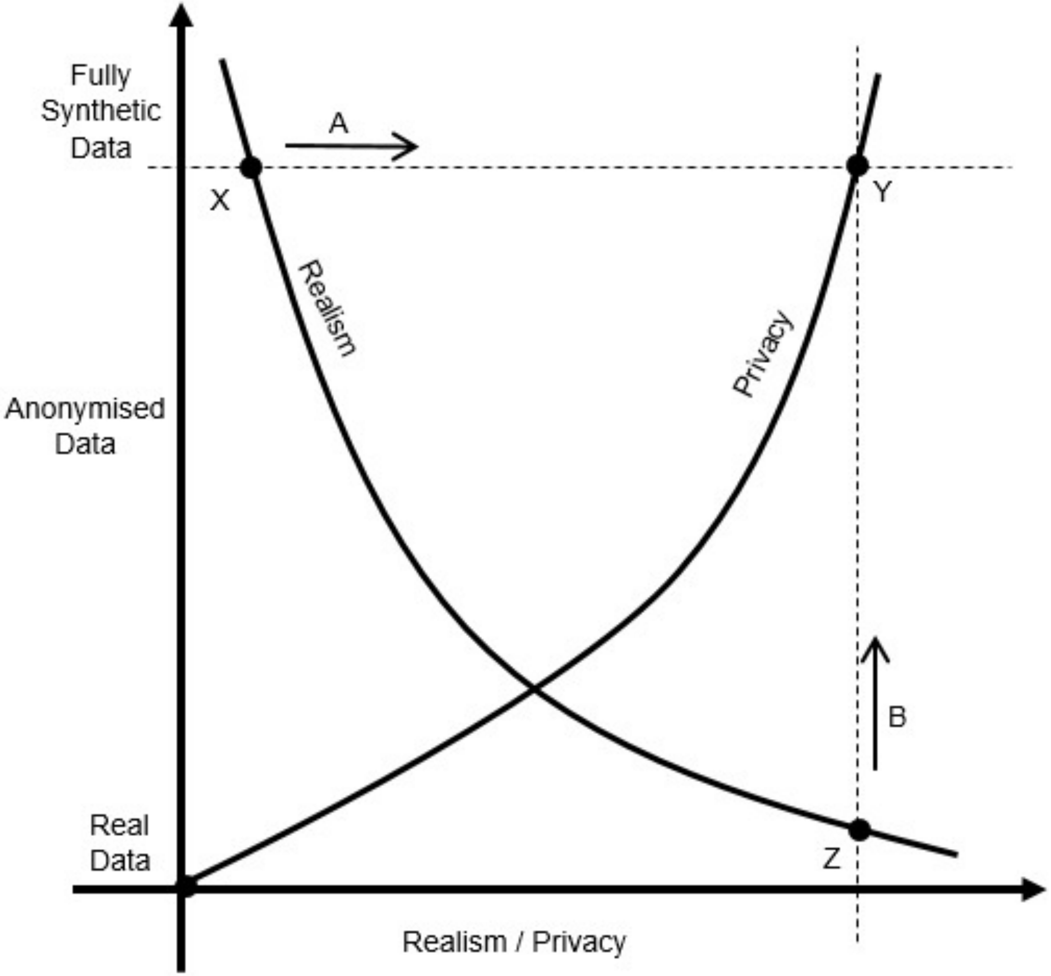
The goal of many SDG projects is…

Creation of a realistic replacement for real data

Realism is seen to bring:

- Greater accuracy
- Reliability
- Effectiveness
- Credibility
- Validity

Synthetic Data Generation



Synthetic Data Generation

Published SDG methods generally failed adherence to the scientific method.

Lacking:

- Complete documentation with full disclosure
- A robust validation method

Given that every author asserts some success in their model…

These are necessary to validate realism

And to justify claims of success

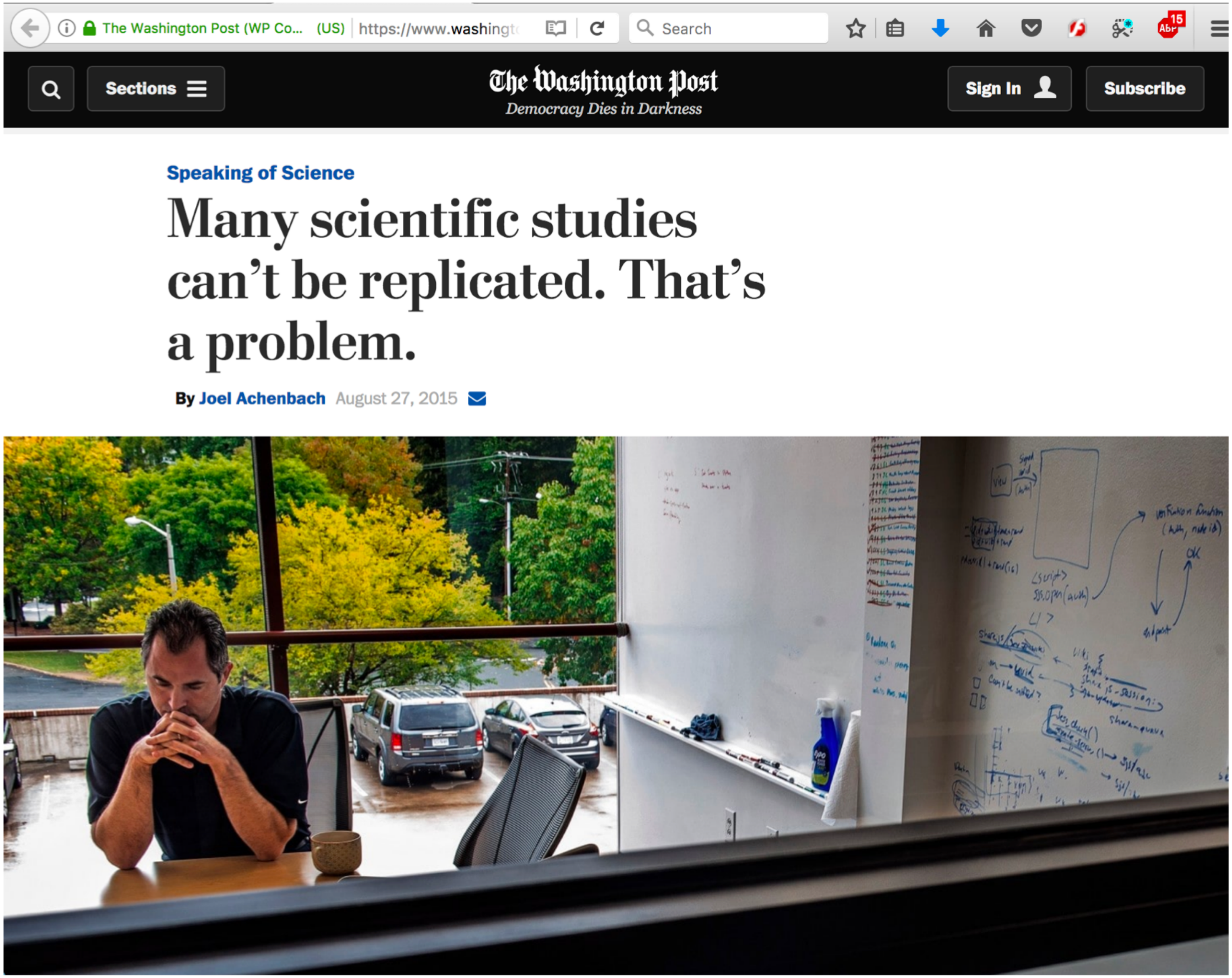
THE SCIENTIFIC METHOD... FOR TEN-YEAR OLDS

DEFINE THE QUESTION
CAN I... WITHOUT ANY HELP... THROW ONE OF THESE ROCKS INTO THE POND?
GATHER INFORMATION AND RESOURCES
GUYS! GUYS! COME CHECK ME OUT!
FORM A HYPOTHESIS
I'MA THROW THIS ROCK SO HARD, YOU WON'T BELIEVE IT.
PERFORM EXPERIMENT AND COLLECT DATA
EEEERGH.
ANALYZE DATA
OH MAN YOU SEE THAT??
DRAW CONCLUSIONS
I ROCK SO FULLY.
PUBLISH RESULTS FOR PEER REVIEW
WOO! SUCKERS! I TOTALLY DID IT! *
RETEST, BY THIRD PARTY
PFFFT. BETCHA I CAN DO IT, TOO.

* THIS IS SURPRISINGLY CLOSE TO HOW REAL SCIENTISTS ACT AT CONFERENCES.

Massey University

Queen Mary
University of London



The Washington Post

Democracy Dies in Darkness

Speaking of Science

Many scientific studies can't be replicated. That's a problem.

By Joel Achenbach August 27, 2015

Massey University

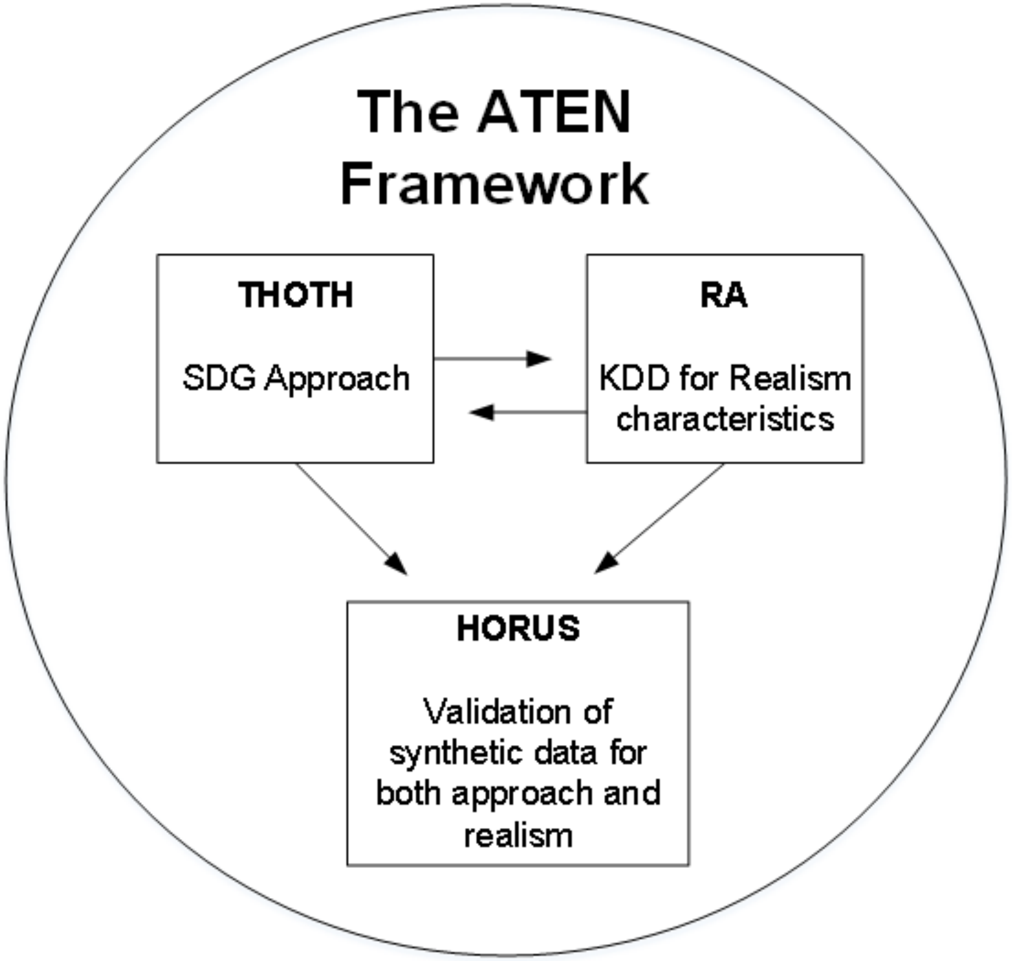
Queen Mary
University of London

RESEARCH PROBLEM

This research sought:

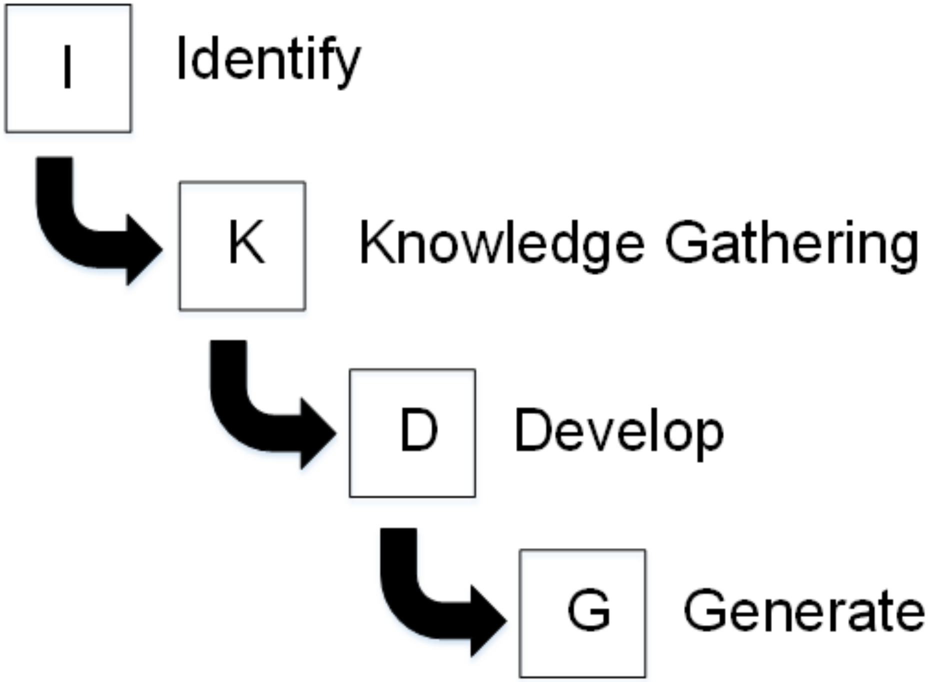
- A generic approach for Synthetic Data Generation (SDG)
- A method for identifying and validating realism in SDG

The ATEN Framework



The ATEN Framework
THOTH
SDG Approach
RA
KDD for Realism characteristics
HORUS
Validation of synthetic data for both approach and realism

THOTH - Generic SDG Approach



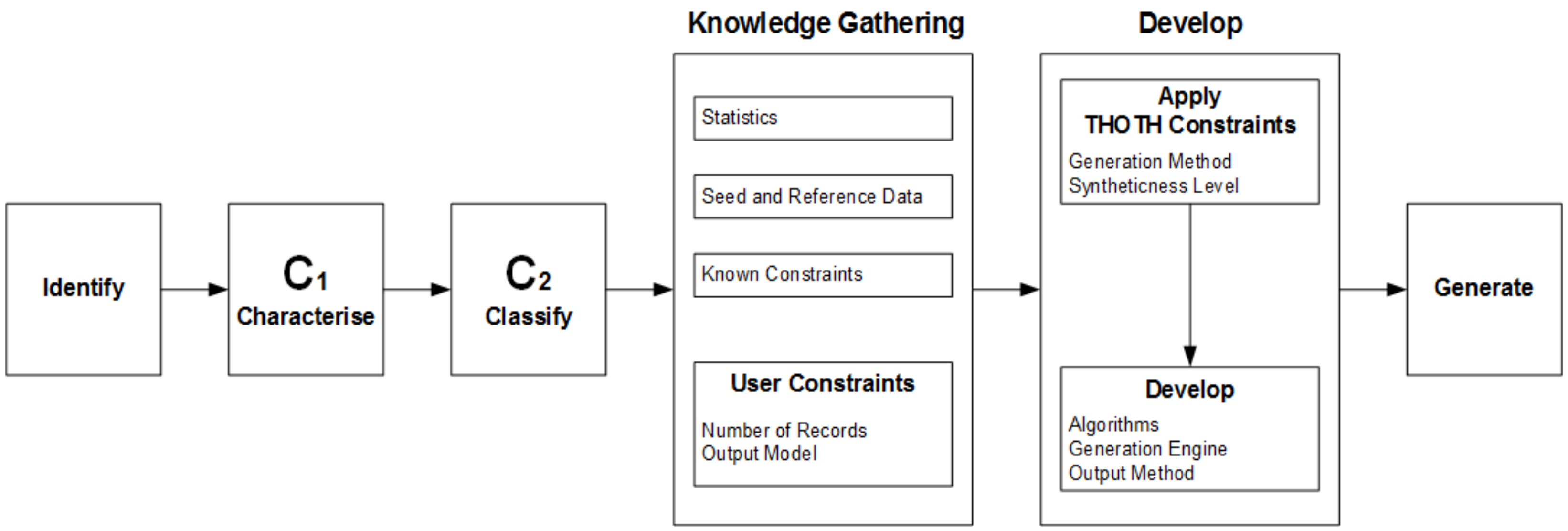
I Identify

K Knowledge Gathering

D Develop

G Generate

THOTH - Enhanced SDG Approach



Identify

C_1 Characterise

C_2 Classify

Knowledge Gathering

Statistics

Seed and Reference Data

Known Constraints

User Constraints

Number of Records
Output Model

Develop

Apply
THO TH Constraints

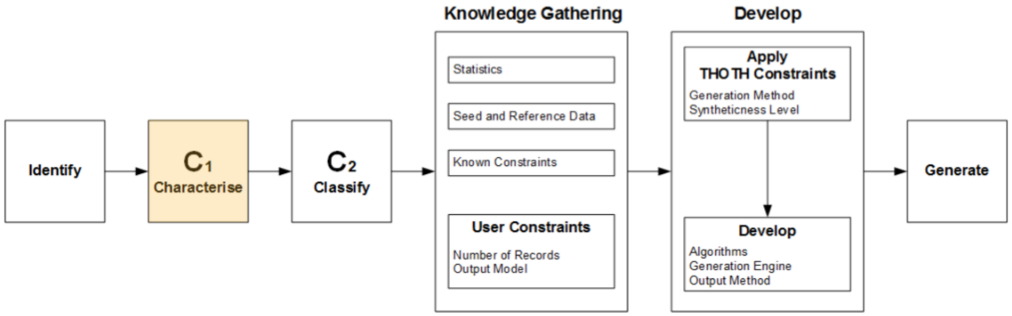
Generation Method
Syntheticness Level

Develop

Algorithms
Generation Engine
Output Method

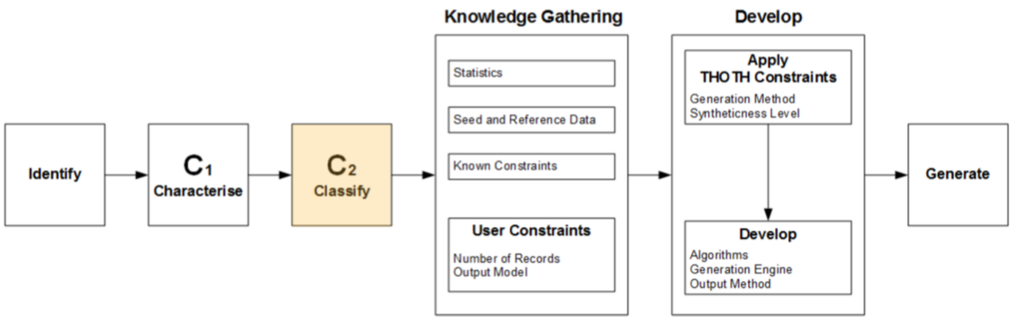
Generate

THOTH - Enhanced SDG Approach



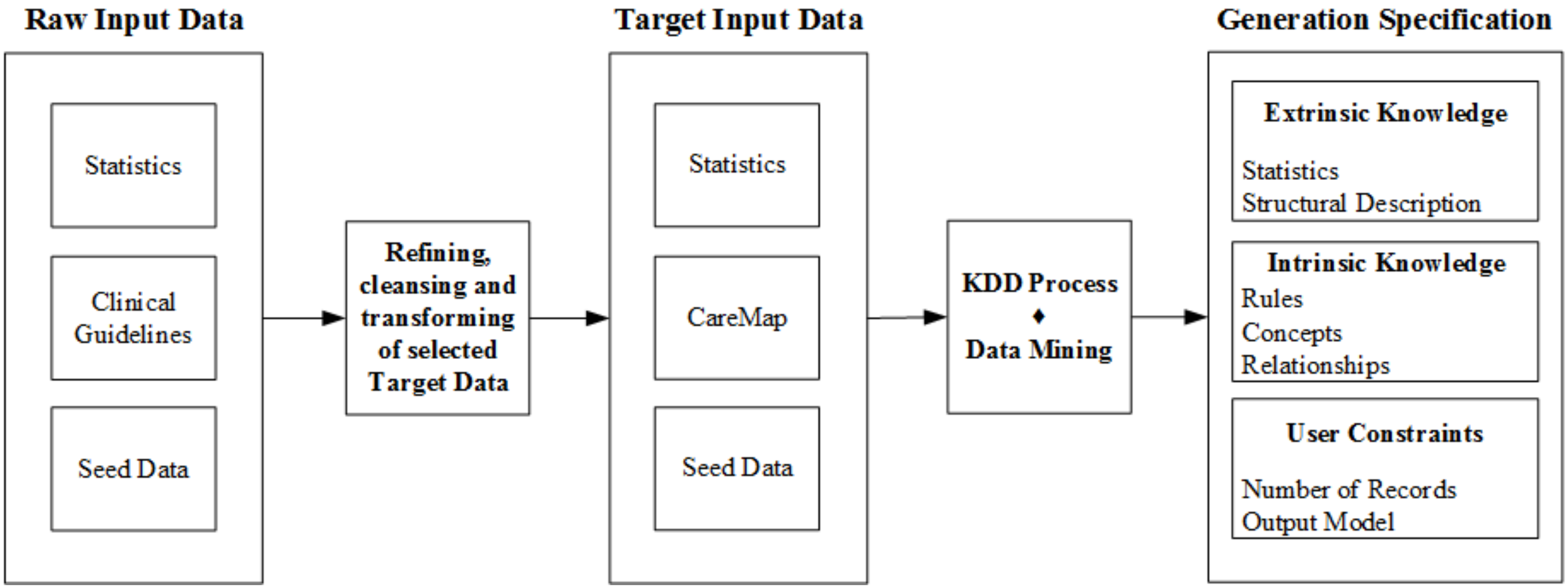
SDG Characterisation	Description
True Synthetic Data	No access to sensitive or confidential data. Uses statistical and probability seed data and expert knowledge. (McLachlan et al, 2016)
Fully Synthetic Data	Requires direct observation of real records to produce synthetic data. (Houkjaer et al, 2006)
Partially Synthetic Data	Unaltered real data is intermixed or aggregated with synthetic data. (Cassa et al, 2004)
Anonymisation-only Data	Some or all PIs have been anonymised (using HIPAA Safe Harbour or similar rules)
Real Data	Real, raw or observed data.

THOTH - Enhanced SDG Approach



SDG Classification Model	Example
Probability Weighted Random Models	Mwogi et al, 2014 Houkjaer et al, 2006 McLachlan et al, 2016
Random Generation Models	Mwogi et al, 2014
Network Generation	Ascoli et al, 2001
Signal and Noise	Whiting et al, 2008
Data Masking	Mouza et al, 2010

RA – Realism in SDG



Raw Input Data

Statistics

Clinical Guidelines

Seed Data

Refining, cleansing and transforming of selected Target Data

Target Input Data

Statistics

CareMap

Seed Data

KDD Process ♦ Data Mining

Generation Specification

Extrinsic Knowledge

Statistics
Structural Description

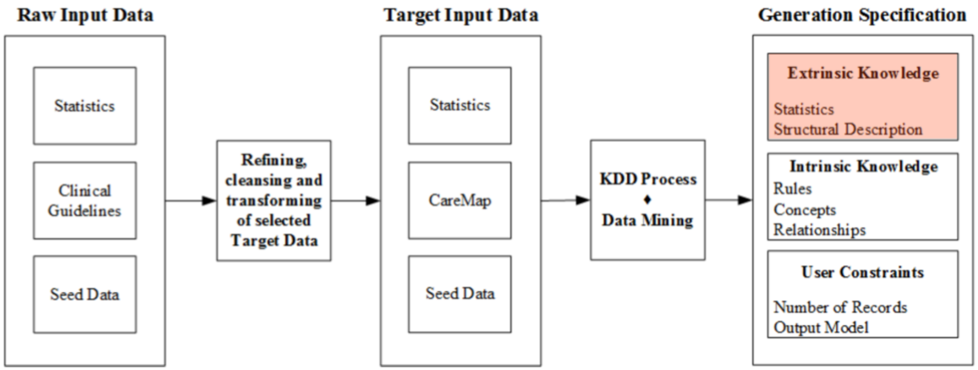
Intrinsic Knowledge

Rules
Concepts
Relationships

User Constraints

Number of Records
Output Model

RA – Realism in SDG



Raw Input Data

Statistics

Clinical Guidelines

Seed Data

Refining, cleansing and transforming of selected Target Data

Target Input Data

Statistics

CareMap

Seed Data

KDD Process ♦ Data Mining

Generation Specification

Extrinsic Knowledge
Statistics
Structural Description

Intrinsic Knowledge
Rules
Concepts
Relationships

User Constraints
Number of Records
Output Model

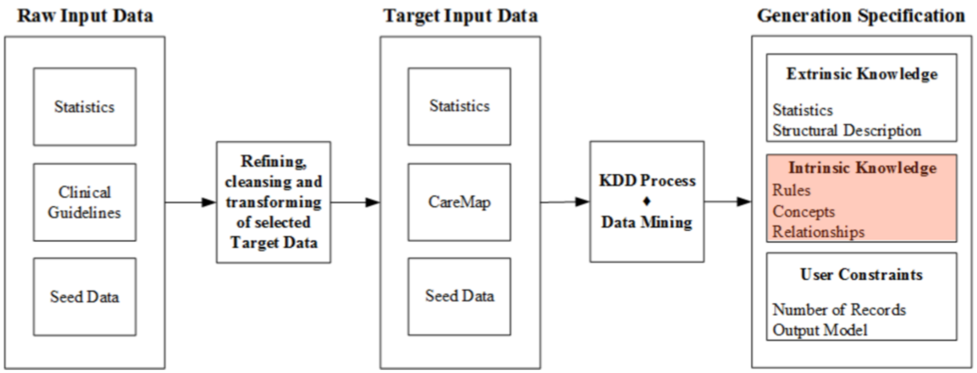
Quantitative

Patient Ethnicity (%)	
European	22.24
Maori	25.13
Pacific Islander	34.30
Asian	16.14
Other	2.11
Not Stated	0.08

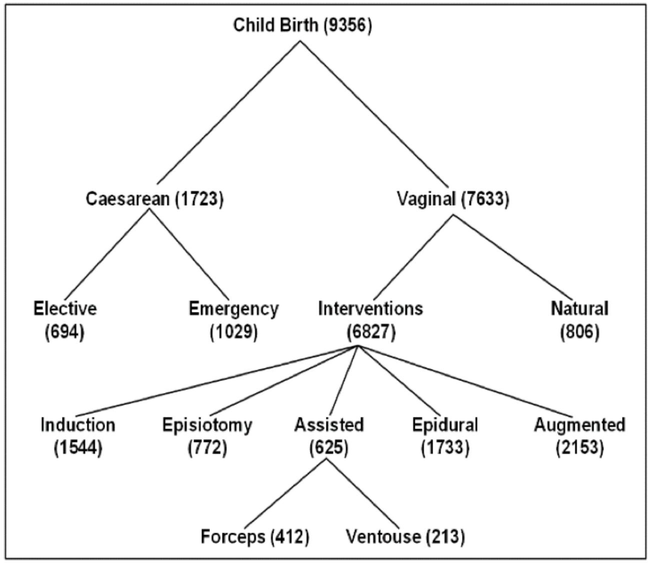
Qualitative

Patient		
PK	patientID	INT
	title	TEXT(10)
	lastName	TEXT(30)
	firstName	TEXT(30)
	dateOfBirth	DATETIME
	gender	CHAR(10)

RA – Realism in SDG



Concept Hierarchies



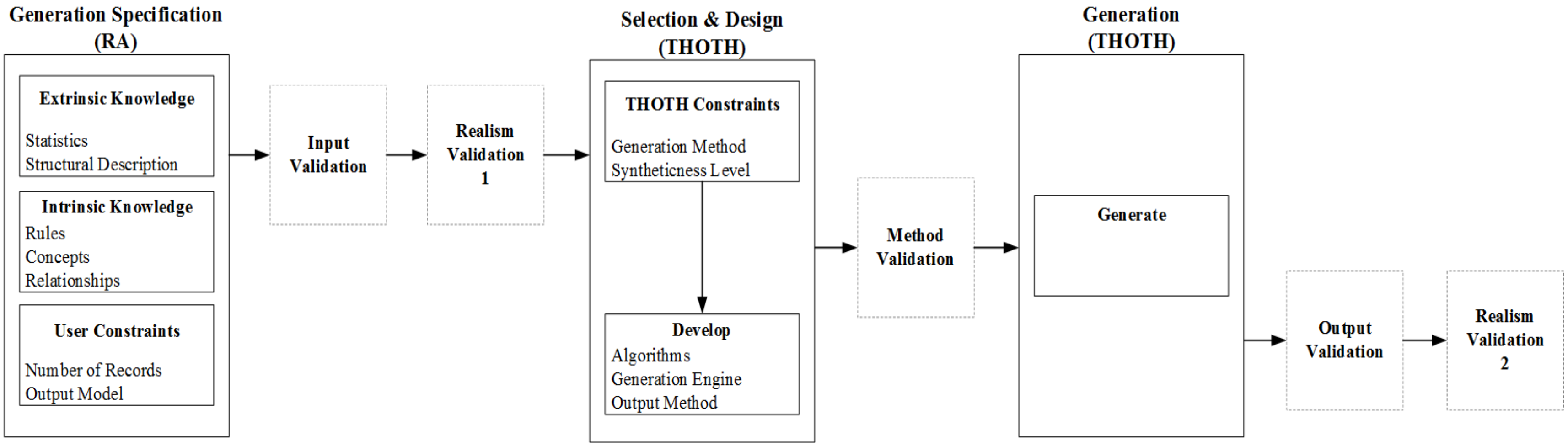
Characteristic Rule

∀χ (midwiferyPatient(x) → ((Sex(x) = female) ^ (Pregnant(x) = Yes) ^ (pregnancyStatus(x) = Low Risk) ^ (fetalHeartMonitoring(x) = Intermittent)))

Classification Rule

∀χ (modeOfDelivery(x) → ((*Multip*(x) = Yes) ^ (*Primip*(x) = No) ^ (*previousDelivery=CSect<2*(x) = No) ^ (*previousDelivery=CSect>=2*(x) = Yes[*d*:100%])))

HORUS – Validation of Realism



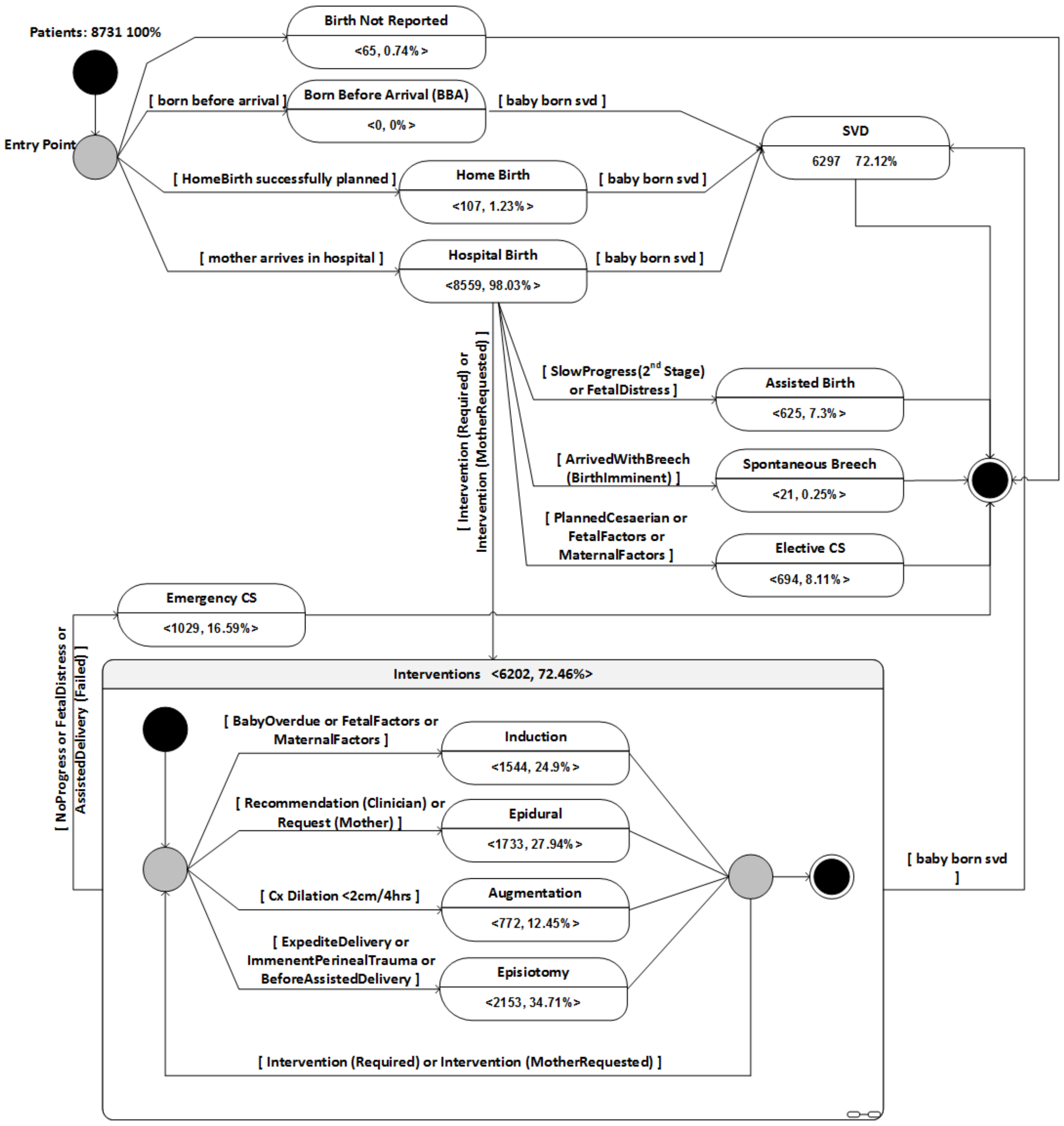
HORUS – Validation of Realism

Step	Activity	Task
1	Input Validation	• Verify each piece of input data or information; • Confirm correctness & validity of input data & information
2	Realism Validation I (RV1)	• Verify concepts & rules derived from the KDD process & health statistical information applied; • Review & test premise & accuracy of each rule to ensure consistency with domain semantics • Tests rules and semantics in real circumstances to eliminate irrelevancy due to interaction with observed data
3	Method Validation	• Review method and compare with others found in literature; • Ensures chosen method is appropriate for generating the synthetic data; and • Verify that the algorithm for the method to be used has been correctly and completely constructed
4	Output Validation	• Establish that output of the SDG model are consistent with observational data; and • Ensure that synthetically generated data conforms to qualitative and quantitative aspects derived during the knowledge discovery phase.
5	Realism Validation II (RV2)	• Perform the same tasks as for Realism Validation I (RV1)

Experiment Design

Labour and Birth

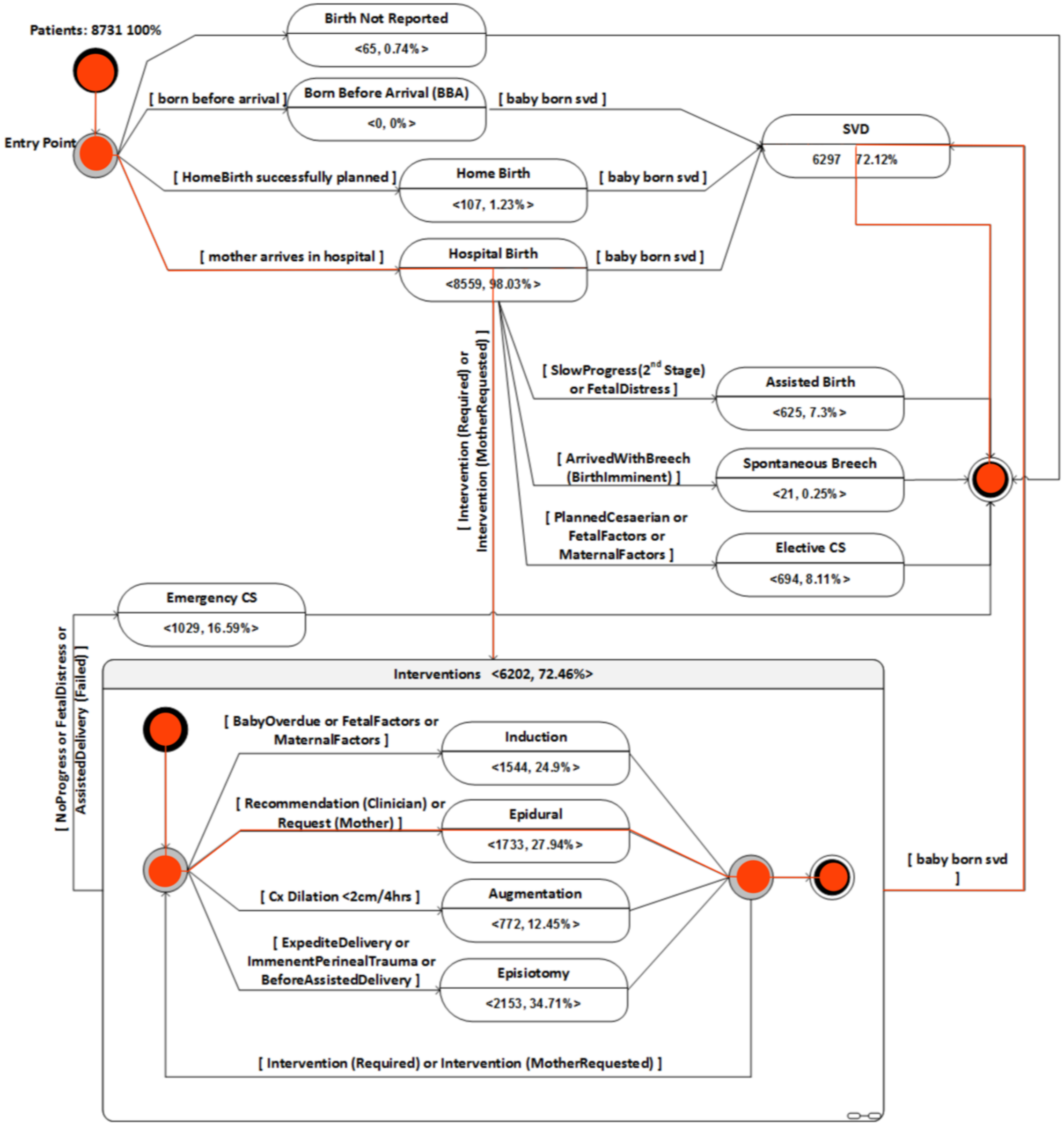
Based on births in the Counties Manukau District Health Board region of NZ for 2014.


Patients: 8731 100%
Entry Point
Birth Not Reported
<65, 0.74%>
[born before arrival]
Born Before Arrival (BBA)
<0, 0%>
[baby born svd]
[HomeBirth successfully planned]
Home Birth
<107, 1.23%>
[baby born svd]
[mother arrives in hospital]
Hospital Birth
<8559, 98.03%>
[baby born svd]
SVD
6297 72.12%
[SlowProgress(2nd Stage) or FetalDistress]
Assisted Birth
<625, 7.3%>
[ArrivedWithBreech (BirthImminent)]
Spontaneous Breech
<21, 0.25%>
[PlannedCesaerian or FetalFactors or MaternalFactors]
Elective CS
<694, 8.11%>
[Intervention (Required) or Intervention (MotherRequested)]
Emergency CS
<1029, 16.59%>
[NoProgress or FetalDistress or AssistedDelivery (Failed)]
Interventions <6202, 72.46%>
[BabyOverdue or FetalFactors or MaternalFactors]
Induction
<1544, 24.9%>
[Recommendation (Clinician) or Request (Mother)]
Epidural
<1733, 27.94%>
[Cx Dilation <2cm/4hrs]
Augmentation
<772, 12.45%>
[ExpediteDelivery or ImmenentPerinealTrauma or BeforeAssistedDelivery]
Episiotomy
<2153, 34.71%>
[Intervention (Required) or Intervention (MotherRequested)]
[baby born svd]

Experiment Design

Labour and Birth

Based on births in the Counties Manukau District Health Board region of NZ for 2014.



Patients: 8731 100%

Entry Point

Birth Not Reported <65, 0.74%>

[born before arrival]

Born Before Arrival (BBA) <0, 0%>

[baby born svd]

[HomeBirth successfully planned]

Home Birth <107, 1.23%>

[baby born svd]

[mother arrives in hospital]

Hospital Birth <8559, 98.03%>

[baby born svd]

SVD 6297 72.12%

[Intervention (Required) or Intervention (MotherRequested)]

[SlowProgress(2nd Stage) or FetalDistress]

Assisted Birth <625, 7.3%>

[ArrivedWithBreech (BirthImminent)]

Spontaneous Breech <21, 0.25%>

[PlannedCesaerian or FetalFactors or MaternalFactors]

Elective CS <694, 8.11%>

Emergency CS <1029, 16.59%>

[NoProgress or FetalDistress or AssistedDelivery (Failed)]

Interventions <6202, 72.46%>

[BabyOverdue or FetalFactors or MaternalFactors]

Induction <1544, 24.9%>

[Recommendation (Clinician) or Request (Mother)]

Epidural <1733, 27.94%>

[Cx Dilation <2cm/4hrs]

Augmentation <772, 12.45%>

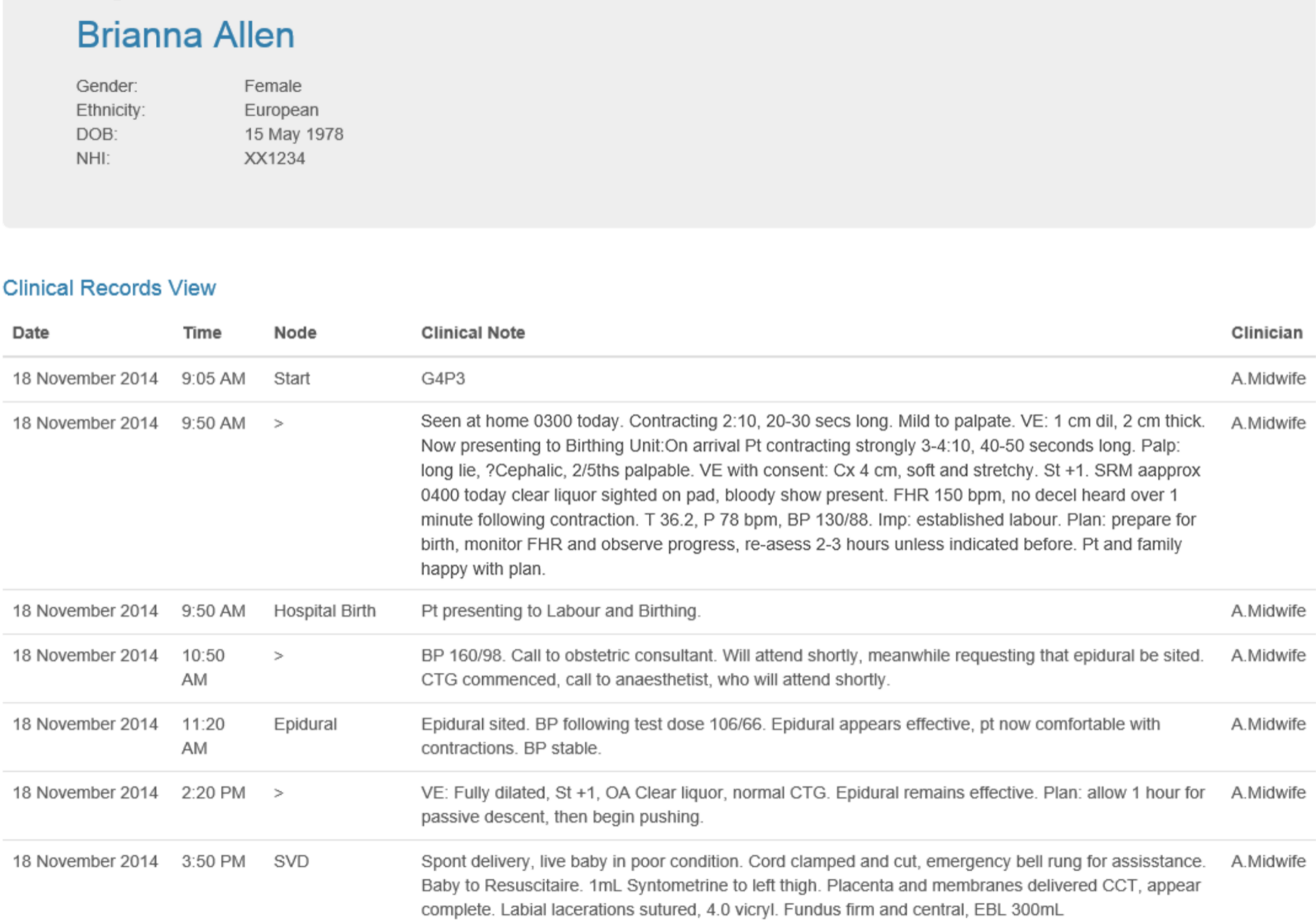
[ExpediteDelivery or ImmenentPerinealTrauma or BeforeAssistedDelivery]

Episiotomy <2153, 34.71%>

[Intervention (Required) or Intervention (MotherRequested)]

[baby born svd]

Experiment Result 1



Brianna Allen

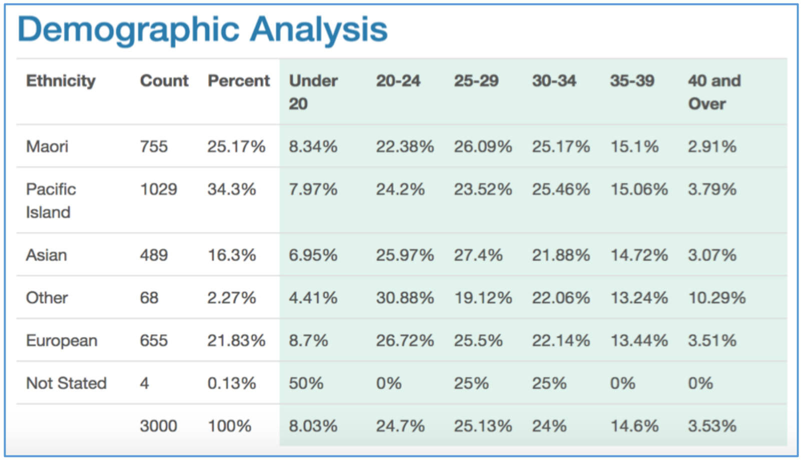
Gender:	Female
Ethnicity:	European
DOB:	15 May 1978
NHI:	XX1234

Clinical Records View

Date	Time	Node	Clinical Note	Clinician
18 November 2014	9:05 AM	Start	G4P3	A.Midwife
18 November 2014	9:50 AM	>	Seen at home 0300 today. Contracting 2:10, 20-30 secs long. Mild to palpate. VE: 1 cm dil, 2 cm thick. Now presenting to Birthing Unit:On arrival Pt contracting strongly 3-4:10, 40-50 seconds long. Palp: long lie, ?Cephalic, 2/5ths palpable. VE with consent: Cx 4 cm, soft and stretchy. St +1. SRM aapprox 0400 today clear liquor sighted on pad, bloody show present. FHR 150 bpm, no decel heard over 1 minute following contraction. T 36.2, P 78 bpm, BP 130/88. Imp: established labour. Plan: prepare for birth, monitor FHR and observe progress, re-asess 2-3 hours unless indicated before. Pt and family happy with plan.	A.Midwife
18 November 2014	9:50 AM	Hospital Birth	Pt presenting to Labour and Birthing.	A.Midwife
18 November 2014	10:50 AM	>	BP 160/98. Call to obstetric consultant. Will attend shortly, meanwhile requesting that epidural be sited. CTG commenced, call to anaesthetist, who will attend shortly.	A.Midwife
18 November 2014	11:20 AM	Epidural	Epidural sited. BP following test dose 106/66. Epidural appears effective, pt now comfortable with contractions. BP stable.	A.Midwife
18 November 2014	2:20 PM	>	VE: Fully dilated, St +1, OA Clear liquor, normal CTG. Epidural remains effective. Plan: allow 1 hour for passive descent, then begin pushing.	A.Midwife
18 November 2014	3:50 PM	SVD	Spont delivery, live baby in poor condition. Cord clamped and cut, emergency bell rung for assisstance. Baby to Resuscitaire. 1mL Syntometrine to left thigh. Placenta and membranes delivered CCT, appear complete. Labial lacerations sutured, 4.0 vicryl. Fundus firm and central, EBL 300mL	A.Midwife

Experiment Result 2

Sample Realistic Synthetic EHR



Demographic Analysis

Ethnicity	Count	Percent	Under 20	20-24	25-29	30-34	35-39	40 and Over
Maori	755	25.17%	8.34%	22.38%	26.09%	25.17%	15.1%	2.91%
Pacific Island	1029	34.3%	7.97%	24.2%	23.52%	25.46%	15.06%	3.79%
Asian	489	16.3%	6.95%	25.97%	27.4%	21.88%	14.72%	3.07%
Other	68	2.27%	4.41%	30.88%	19.12%	22.06%	13.24%	10.29%
European	655	21.83%	8.7%	26.72%	25.5%	22.14%	13.44%	3.51%
Not Stated	4	0.13%	50%	0%	25%	25%	0%	0%
	3000	100%	8.03%	24.7%	25.13%	24%	14.6%	3.53%

CoMSER Dashboard: Demographic Analysis Table

Ethnicity	Statistical %	CoMSER SDG
Maori	25.13	25.17
Pacific Islander	34.30	34.30
Asian	16.14	16.30
Other	02.11	2.27
European	22.24	21.83
Not Stated	00.08	0.13

Ethnicity Statistical Comparison

Age Range	Statistical %	CoMSER SDG
Under 20	8.26	8.03
20-24	22.93	24.70
25-29	26.74	25.13
30-34	23.96	24.00
35-39	14.58	14.60
40 and Over	3.53	3.53

Age Statistical Comparison

SUMMARY AND CONCLUSIONS

- Attaining realism in synthetically generated datasets is challenging

- SDG authors claim success in generating realistic synthetic data, yet:
  - Few identify the elements of realism from real data that they seek to recreate;
  - Fewer still present validation of realism in the resulting synthetic data.

- ATEN provides a complete and generic method for SDG that;
  - Identifies the level of synthetic-ness required;
  - Helps to classify the generation method required;
  - Identifies necessary realistic elements in real data, and;
  - Allows for validation of their existence in the resulting data.

And, most importantly: **ATEN supports claims of success in SDG**

Questions

UNIVERSITY OF MONTANA

MITRE

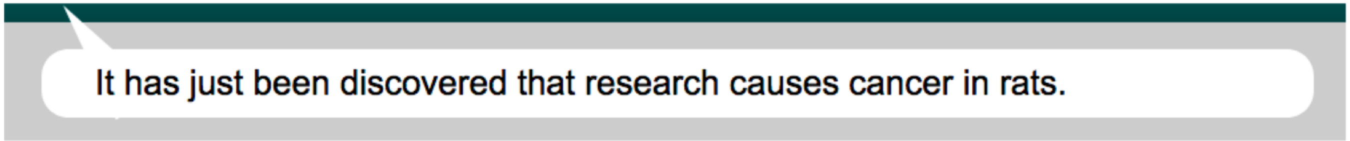
PAMBAYESIAN

EPSRC
Engineering and Physical Sciences
Research Council

Massey University

Queen Mary
University of London

A Final Observation…



It has just been discovered that research causes cancer in rats.