Synthetix Health

Stockholm, Sweden

Thesis programme 2027 · KTH · KI · SU

Check our work.

Master's and PhD projects on synthetic clinical data that has to prove itself, with a question worth publishing in each.

Who we are

Synthetix Health is an early-stage Stockholm company building certified synthetic clinical data: artificial patient cohorts produced under formal ε-differential privacy. Each release comes with a certificate that states its privacy budget, what it is fit for, and what it is not fit for.

Why theses

A privacy guarantee or a utility claim is only worth something once someone independent has tried to break it. These projects test our releases at arm's length and build the methods the field needs. Our first clinical focus is diabetes and cardiovascular care.

How it works

- A Synthetix supervisor plus your academic supervisor
- Public seed data from day one; real records only inside approved secure environments
- A confidentiality and IP agreement before work begins
- Negative results are publishable

Topics

	Project	Field	Suits
M1	**Red-teaming synthetic clinical data**	Privacy · Security	KTH (Cybersecurity, ML) · SU (DSV)
M2	**Same data, same conclusions?**	Biostatistics · Epidemiology	KI (Epidemiology, Biostatistics & Data Science) · SU (Statistics)
M3	**Utility that matches the use**	Statistics · Machine learning	KTH (ML, Data Science) · SU (Statistics, CS)
M4	**Interoperable synthetic data**	Health informatics	KI / SU (Health Informatics) · KTH (Medical Engineering)
M5	**Who pays for privacy?**	Statistics · Public health	KI (Public Health, Epidemiology) · KTH (ML)
M6	**What makes a release adequate?**	Law · Policy	SU (Law & IT, European law) · KI (Health policy)
M7	**Validation inside the firewall**	Health informatics · Implementation	KI (LIME, Health Informatics) · KTH (Health systems)
M8	**Synthetic data across drug development**	Regulatory science · Bioentrepreneurship	KI (Bioentrepreneurship, Regulatory science) · SU
P1	**Auditable privacy for clinical synthetic data**	Privacy · Security	KTH · SU · co-supervised
P2	**Fit-for-purpose evidence for synthetic data**	Epidemiology · Regulatory science	KI · co-supervised

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M1

Red-teaming synthetic clinical data

Membership-inference, reconstruction and linkage attacks against certified synthetic releases.

Master's · 30 credits | Privacy · Security | Suits: KTH (Cybersecurity, ML) · SU (DSV)

The question

A synthetic release produced under ε-differential privacy comes with a mathematical ceiling on what any attacker can learn about a real patient. How close do real, well-resourced attacks get to that ceiling, and where does the gap come from?

Why it matters

Formal guarantees are only persuasive to a hospital or an ethics board when someone independent has tried hard to break them. Published attacks on synthetic data show that similarity scores can hide real leakage. An open, reproducible attack harness is the evidence the field is missing.

What you would do

01 Survey and implement the main attack families from the literature: shadow-model membership inference, distance-based attacks, attribute inference and record linkage.

02 Run them against public-seed releases (CMS DE-SynPUF) at several privacy budgets, with a clear threat model for each attack.

03 Compare measured attack advantage against the theoretical bound printed on the release certificate.

04 Where a partner's secure environment and ethics approval allow it, repeat the strongest attacks on releases made from real records, without data leaving the environment.

What you would produce

- An open-source attack harness with documented threat models
- A results report: where attacks come close to the bound, and where they do not
- A publishable paper; negative results count

Good fit if you have

Python, ML fundamentals, some probability. Interest in security or cryptography. Prior work with privacy attacks is a plus, not a requirement.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M2

Same data, same conclusions?

Replicating published diabetes and cardiovascular analyses on synthetic cohorts and measuring inference agreement.

Master's · 30 credits | Biostatistics · Epidemiology | Suits: KI (Epidemiology, Biostatistics & Data Science) · SU (Statistics)

The question

When a researcher runs a standard cardiometabolic analysis, such as diabetes outcomes or heart-failure survival, on a synthetic cohort instead of the real one, do they reach the same conclusions: the same direction of effect, overlapping confidence intervals, the same curves?

Why it matters

The point of synthetic data in research is that inferences carry over. Most evaluations report distance metrics instead of asking whether the scientific answer changes. A clear protocol for inference agreement is what a research group would accept before trusting a synthetic cohort. We start in diabetes and cardiovascular care because the conditions are common and the analyses are well established.

What you would do

01 Select three to five published diabetes and cardiovascular analyses that can be reproduced on public data: risk-factor regression, Kaplan–Meier and Cox survival analysis, comorbidity prevalence.

02 Re-run each analysis on the real seed data and on synthetic releases at several privacy budgets.

03 Measure agreement: effect-size difference, confidence-interval overlap, sign and significance agreement, curve distance.

04 Propose a compact inference-agreement report that could sit on a release certificate.

What you would produce

- A reproducibility protocol other groups can reuse
- Agreement results across analyses and budgets
- A short reporting template for inference agreement

Good fit if you have

R or Python, regression and survival analysis, careful reporting. Epidemiology or biostatistics background.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis
All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M3

Utility that matches the use

Which fidelity measures actually predict success on which downstream clinical tasks.

Master's · 30 credits | Statistics · Machine learning | Suits: KTH (ML, Data Science) · SU (Statistics, CS)

The question

Synthetic data is scored with dozens of fidelity metrics. Which of them predict how well a downstream task performs, and does the answer change between software testing, cohort sizing and model training?

Why it matters

A single utility score invites the wrong reading. Buyers and access bodies need to know which evidence counts for which purpose. Mapping metrics to tasks turns fitness for purpose from a slogan into something measurable.

What you would do

01 Catalogue fidelity metrics in common use: marginal divergences, correlation structure, propensity scores, train-synthetic-test-real performance.

02 Define a set of downstream tasks of different difficulty on public seed data.

03 Measure metric scores and task performance across many synthetic releases and look for which metrics are predictive, and where.

04 Recommend a minimal, purpose-specific evidence set.

What you would produce

- A metric-to-task predictiveness study
- A recommended evidence set per purpose
- Open evaluation code

Good fit if you have

Python, statistics, experimental design. Comfortable running many experiments and reasoning about them.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis
All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M4

Interoperable synthetic data

Synthetic cohorts that stay valid against FHIR and OMOP standards.

Master's · 30 credits | Health informatics | Suits: KI / SU (Health Informatics) · KTH (Medical Engineering)

The question

A synthetic cohort is only usable if it loads into the systems researchers already use. How often do synthetic records violate FHIR profiles, OMOP CDM conventions and basic clinical plausibility, and how should that be tested and reported?

Why it matters

Hospitals and research platforms run on standards. Structural errors in synthetic data waste analysts' time and erode trust faster than statistical gaps. A standard conformance suite is a practical, reusable contribution.

What you would do

01 Assemble a conformance suite: FHIR profile validation, OMOP data-quality checks (e.g. OHDSI tooling), referential integrity and clinical plausibility rules.

02 Apply it to synthetic releases and to their public seed data as a baseline.

03 Classify the violations found and how they affect downstream use.

04 Propose a conformance section for a release certificate.

What you would produce

- An open conformance test suite
- A violation taxonomy with examples
- A reporting format for interoperability

Good fit if you have

Health data standards (FHIR, OMOP) or willingness to learn them quickly; SQL and Python.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M5

Who pays for privacy?

Measuring and reporting subgroup-level utility in private synthetic data for diabetes and cardiovascular patients.

Master's · 30 credits | Statistics · Public health | Suits: KI (Public Health, Epidemiology) · KTH (ML)

The question

Under a fixed privacy budget, does synthetic data serve some cardiometabolic patients worse than others, such as younger people with type 2 diabetes, women with heart disease, or older patients with several conditions? How should that be measured and shown to the people relying on the data?

Why it matters

Published work shows that privacy noise can land hardest on small subgroups, and most medical studies do not report subgroup results at all. Regulators and the EU AI Act increasingly ask for subgroup evidence. Measuring it honestly is a design requirement for release documentation.

What you would do

01 Define subgroups among diabetes and cardiovascular patients in public seed data (age, sex, region, comorbidity) and a set of subgroup-level fidelity and inference measures.

02 Measure subgroup gaps across synthetic releases at several privacy budgets.

03 Relate the gaps to subgroup size and to the downstream tasks affected.

04 Propose a subgroup section for a release certificate.

What you would produce

- A subgroup-utility benchmark on public data
- Analysis of which groups lose most, and when
- A reporting template for subgroup results

Good fit if you have

Statistics, R or Python, interest in equity and fairness in health research.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M6

What makes a release adequate?

Criteria a Health Data Access Body could apply to synthetic outputs under the EHDS, GDPR and the AI Act.

Master's · 30 credits | Law · Policy | Suits: SU (Law & IT, European law) · KI (Health policy)

The question

When is a synthetic dataset anonymous enough, and documented well enough, to be released under the European Health Data Space? What should an access body be able to check?

Why it matters

EHDS access bodies will have to approve or refuse synthetic outputs, yet there is no agreed standard for what adequate means. GDPR anonymity, the AI Act's data-governance rules and EHDS secondary-use rules each pull in a different direction. A clear set of criteria is useful to regulators and to every producer.

What you would do

01 Map the relevant provisions: GDPR anonymisation, EHDS secondary use, AI Act data governance, relevant EDPB guidance.

02 Analyse how formal privacy guarantees and empirical risk assessments fit those provisions.

03 Compare approaches in two or three member states or authorities.

04 Propose checkable adequacy criteria, including what a release should state it must not be used for.

What you would produce

- A doctrinal and comparative analysis
- Proposed adequacy criteria for access bodies
- A policy brief suitable for publication

Good fit if you have

EU law, data protection law. Comfort reading technical summaries; no programming needed.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M7

Validation inside the firewall

A protocol for evaluating synthetic data in a hospital's secure environment, where raw records never leave.

Master's · 30 credits | Health informatics · Implementation | Suits: KI (LIME, Health Informatics) · KTH (Health systems)

The question

What does it take for a diabetes or cardiology clinic to host a synthetic-data validation, so that raw patient records stay inside its environment and only synthetic data and evaluation results come out?

Why it matters

The strongest evidence for synthetic data comes from real clinical records. We start in cardiometabolic care, at a diabetes or cardiology clinic, with oncology as a later stage. Hospitals will only take part if governance is clear: ethics approval, DPIA, roles, access controls and exit criteria. A tested protocol lowers the barrier for the first sites.

What you would do

01 Review existing secure-processing models (e.g. trusted research environments, MONA-style remote access) and Swedish requirements.

02 Interview data stewards, ethics and IT security staff about their conditions for participating.

03 Draft a step-by-step validation protocol: approvals, roles, what runs inside, what may leave, stop criteria.

04 Pilot the protocol end-to-end on public cardiometabolic data in a simulated environment.

What you would produce

- A governance and validation protocol hospitals can review
- Interview findings on adoption barriers
- Checklists and templates for a first site

Good fit if you have

Health informatics or health systems background; interviewing and qualitative analysis; basic technical literacy.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

M8

Synthetic data across drug development

What evidence counts at each stage, followed through one cardiometabolic therapy area from discovery to regulatory submission.

Master's · 30 credits | Regulatory science · Bioentrepreneurship | Suits: KI (Bioentrepreneurship, Regulatory science) · SU

The question

Where in the drug-development pipeline for cardiometabolic disease can certified synthetic data be used today, and what evidence would a pharma team, an ethics board or a regulator need at each stage?

Why it matters

Synthetic data is discussed for discovery, trial feasibility and synthetic control arms, but expectations differ sharply by stage. EMA and FDA statements give partial guidance. A clear stage-by-stage evidence map helps companies and research institutes decide where to start.

What you would do

01 Pick one cardiometabolic therapy area, such as type 2 diabetes or heart failure, and map its stages: target discovery, cohort and feasibility analysis, trial design, external comparators, post-market.

02 Review EMA, FDA and Nordic guidance and published use cases at each stage.

03 Interview people in Nordic pharma, CROs and research institutes about what they would accept.

04 Build an evidence ladder: which uses are ready, which need more validation, which are not appropriate.

What you would produce

- A stage-by-stage evidence map
- Interview insights from industry
- Recommendations for early adoption pathways

Good fit if you have

Life-science background; interest in regulation and business; interviewing and synthesis.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

P1

Auditable privacy for clinical synthetic data

Closing the gap between formal guarantees and what attacks achieve in practice.

PhD direction | Privacy · Security | Suits: KTH · SU · co-supervised

The question

Can we build audit methods that let an independent party verify, with statistical confidence, that a synthetic clinical release behaves as its stated privacy guarantee says, without seeing the producer's internals?

Why it matters

Privacy auditing is an active research area, but clinical synthetic data has its own structure, such as linked tables and coded clinical events. Third-party verifiability is what turns a vendor claim into a trust standard.

What you would do

01 Develop and evaluate black-box audit methods for synthetic tabular and relational clinical data.

02 Study how tight empirical lower bounds on privacy loss can be made for realistic data.

03 Validate on public data and, where approved, inside partner secure environments.

04 Publish open audit tooling.

What you would produce

- A research programme of three to four papers
- Open audit tooling
- Collaboration across academia, hospitals and industry

Good fit if you have

Strong background in ML, statistics or security; master's degree in a relevant field.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Start by agreement. We are an early-stage company and cannot fund a PhD position alone; this is a direction for a co-supervised project if a research group and funding (e.g. an industrial doctorate) align.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis

All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.

P2

Fit-for-purpose evidence for synthetic data

From inference reproducibility to regulatory acceptance in clinical research.

PhD direction | Epidemiology · Regulatory science | Suits: KI · co-supervised

The question

What body of evidence should be required before synthetic clinical data is used for a given research purpose, and how can that evidence be standardised so registries, access bodies and regulators accept it?

Why it matters

There is no reporting guideline for synthetic health data comparable to those for trials or observational studies. Sweden's registers and secure environments make it an unusually good place to build and test one.

What you would do

01 Systematically review evaluation practice and existing reporting guidelines.

02 Run reproducibility studies on public and, where approved, register-based data.

03 Develop and test a reporting guideline through consensus methods with clinicians, statisticians and regulators.

04 Study adoption in real research projects.

What you would produce

- A proposed reporting guideline
- A series of reproducibility studies
- Evidence on adoption and acceptance

Good fit if you have

Epidemiology or biostatistics master's; interest in research methods and regulation.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Start by agreement. We are an early-stage company and cannot fund a PhD position alone; this is a direction for a co-supervised project if a research group and funding (e.g. an industrial doctorate) align.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis
All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.