NIPH
Norwegian Institute of Public Health

REPORT

2023

Strategy proposal for further implementation of machine learning in the Cluster of Reviews and Health Technology Assessments

Utgitt av	Folkehelseinstituttet Område for helsetjenester
Tittel	Strategiforslag for videre implementering av maskinlæring i klynge for vurdering av tiltak
English title	Strategy proposal for further implementation of machine learning in the Cluster of Reviews and Health Technology Assessments
Ansvarlig	Camilla Stoltenberg, direktør
Forfattere	Tiril C. Borge, Folkehelseinstituttet Hans Bugge Bergsund, Folkehelseinstituttet Jan P.W. Himmels, Folkehelseinstituttet Jose F Meneses-Echavez, Folkehelseinstituttet Christopher Rose, Folkehelseinstituttet Ashley Elizabeth Muller, prosjektleder, Folkehelseinstituttet
ISBN	978-82-8406-358-4
Publikasjonstype	Report (Rapport)
Antall sider	18
Oppdragsgiver	Folkehelseinstituttet
Emneord(MeSH)	biomedical; technological assessment, health; unsupervised machine learning; supervised machine learning; deep learning
Sitering	Borge TC, Bergsund HB, Himmels JPW, Meneses-Echavez JF, Rose C, Muller AE, (2023). "Strategiforslag for videre implementering av maskinlæring i klynge for vurdering av tiltak" [Strategy proposal for further implementation of machine learning in the Cluster of Reviews and Health Technology Assessments]. Oslo: Folkehelseinstituttet.

Contents

Key Messages

Since 2020 the Cluster for Reviews and Health Technology Assessments (HTV) has implemented machine learning (ML) in the production of evidence syntheses and health technology assessments. This was due to a need and desire to streamline such research processes, as the gold standard methods are resource-intensive, making current practices unsustainable. ML can automate complex, repetitive tasks in evidence synthesis processes, thus reducing resource requirements.

HTV established two dedicated teams for the implementation of ML: ML 1.0 (2020-2021) and ML 2.0 (2021-2022). The ML teams have been very successful with this work: they have documented workload savings, as well as established themselves as implementation leaders in the field. The experiences from the ML work in HTV since 2020 form the basis for this report, which describes ML 2.0's strategy proposal for how HTV should implement further work with ML in evidence synthesis processes:

- **ML 3.0 maintain focus on exploration**, to identify new functions and applications.
- **Existing HTV expertise structures** (e.g. *undervisningslaget*) take responsibility for building employee capacity to use established functions.
- **Team 3.0 retains responsibilty for building capacity** of novel functions and applications, and all such training must be **scalable**, e.g. asynchronous and interactive online trainings.

A final suggestion is for HTV to collect existing and new **development and innovation activities into one portfolio in the cluster**. This portfolio could include ML Team 3.0, as well as other types of projects, teams, and activities related to automation, digitalization, and process change. An area of co-generative learning would be created, as well as an incubator for funding applications.

Title:
Strategy proposal for further implementation of machine learning in the Cluster of Reviews and Health Technology Assessments

---

Publisher:
The Cluster of Reviews and Health Technology Assessments, Division for Health Services established the ML Team 2.0 innovation team as a response to the needs expressed by the Norwegian Institute of Public Health

---

Type of publication:
Report

Hovedbudskap

Klynge for vurdering av tiltak (HTV) har siden 2020 jobbet med å innføre maskinlæring (ML) i utarbeidelsen av kunnskapsoppsummeringer og metodevurderinger. Behovet og ønsket var effektivisering av slike forskningsleveranser, da gullstandardmetodene er resurskrevende, hvilket gjør dagens praksis lite bærekraftig. ML kan automatisere komplekse, repetitive oppgaver i kunnskapsoppsummeringsprosessen, og dermed redusere ressursbehovet.

HTV etablerte to lag (team) dedikert til innføring av ML: ML 1.0 (2020-2021) og ML 2.0 (2021-2022). ML-lagene har hatt stor suksess med dette arbeidet: de har blant annet kunnet dokumentere arbeidsbesparelser og har etablert seg som en implementeringsleder på feltet. Erfaringene fra ML arbeidet siden 2020 danner grunnlaget for denne rapporten, som beskriver ML 2.0 sitt strategiforslag til hvordan HTV bør implementere det videre arbeidet med ML i kunnskapsoppsummeringsprosesser:

- **ML 3.0 holder fokus på utforskning** for å blant annet identifisere nye funksjoner og applikasjoner
- **Eksisterende HTV kompetansebyggingsstrukturer** tar over ansvar for å bygge medarbeidernes kompetanse til å bruke etablerte ML-funksjoner.
- **ML 3.0 beholder ansvaret for å bygge kompetanse** til nye funksjoner og applikasjoner, og all slik opplæring må være **skalerbar**, f.eks. ved hjelp av interaktive nettbaserte opplæringsmoduler.

En siste anbefaling er at HTV samler eksisterende **utviklings- og innovasjonsaktiviteter i en portefølje i klyngen**. Den kan inkludere ML 3.0, så vel som andre typer prosjekter, teams og aktiviteter knyttet til automatisering, digitalisering og prosessendring. Et område for samskapt læring vil bli opprettet, samt en inkubator for søknader om finansiering.

Tittel
Strategiforslag for videre implementering av maskinlæring i klynge for vurdering av tiltak

---

Hvem står bak denne publikasjonen?
Folkehelseinstitutet utførte studien basert på et initiativ fra klynge for vurdering av tiltak, område for helsetjenster i FHI

Type publikasjon:
Rapport

Preface

This report presents strategy suggestions for the next iteration of the machine learning team, "ML 3.0". The current team, ML 2.0, has crafted these suggestions based on our reflections of our wins, losses, and learning in the period 2021-2022.

Financing
The work was self-initiated and financed by the Cluster for Reviews and Health Technology Assessments, Division for Health Services at the NIPH.

With appreciation
The current team's learning and strategizing are due not only to the dedication of its members, past and present, but also to HTV management's investment and vocal support. There have also been numerous colleagues who have provided support, feedback, ideas, and opportunities. This strategy document has benefited greatly from the experimentation and involvement of Alexandra Poulsson, Marit Austeng's networks and encouragement to think larger, Knut Børtnes' coaching around change communication, and Ragnhild Valen's enthusiasm and idea-sharing. Outside of NIPH, James Thomas' mentoring and his team at EPPI Centre have continued to be instrumental to our understanding of ML and its potential to provide the most valuable evidence synthesis products to our commissioners.

Conflicts of interest
All authors declare they have no conflicts of interest.

Kåre Birger Hagen	Rigmor C Berg	Ashley E. Muller
Research director	*Department director*	*Project leader*

Machine learning as an exploratory strategic innovation

As technology changes, so does NIPH's Cluster for Reviews and Health Technology Assessments' market for evidence synthesis. The types of products we provide today, and that our commissioners choose us over alternatives, is not a given, and we need to continually change to make sure we are providing value. The cluster must therefore engage in strategic innovation, i.e. changing its working model to ensure a sustainable competitive advantage over other organizations (1). Strategic innovation entails direct changes to products, and it can entail changes to the working processes, commissioner bases, relationships with our target audiences. Ultimately, strategic innovation should help our cluster provide more value than any other organization.

HTV has two competing needs related to strategic innovation: first, it must build its ***expertise*** and enhance its capabilities – it must get better at what it does, and ideally, be the best environment for evidence synthesis and health technology assessments. The indicator of capability enhancement is improvement. The second need is not to get better at playing the game, but to change the rules of the game – to ***explore*** what else we can do/produce/learn, that lies outside the path of status quo improvement. Exploration is to move off the expected trajectory, and without knowing where one will land. Strategic innovation is therefore described as a paradox, as these competing needs require managerial decision-making about resources, skill sets, and directions.

The cluster has well-developed and high-quality procedures for capacity-building of employees, such as internal and external teaching to upskill employees, strategically hiring experts in new methods, using inter-divisional networks as learning arenas, and pedagogic quality control procedures and standards. NIPH is a center of excellence within evidence synthesis because of the cluster's investment in experts and dedication to methodological gold standards. These gold standards make capacity-building smoother, as they represent a broad agreement of what "the best" looks like.

ML Teams 1.0 and 2.0

The first machine learning (ML) team, **ML 1.0**, was tasked to focus on ML exploration with the mandate to i) test and document the advantages and disadvantages of using machine learning in the review process, and ii) building ML capacity in the klynge. There were not yet standards

of ML that could be improved upon, or consensus around what should be taught, or minimum levels of capacity that needed to be enhanced.

The team grew in the second iteration, ***ML 2.0***, with a larger but more specific mandate: to identify innovative ML functions and applications, evaluate these for fit-for-purpose, and build capacity within our division to use the "successful" functions independently of the ML team. The new team represented a focus on exploration compared with the rest of the cluster, but there was also a focus on exploration as well as expertise *within* the team. Just as HTV as a whole has a responsibility to constantly produce the best products possible, i.e. improving its evidence synthesis expertise, ML 2.0 now also had to utilize and enhance its ML-related capabilities. Adding capacity-building to ML 2.0's mandate meant upskilling the team and the cluster systematically, refining procedures, and creating higher standards for understanding and use.

ML Team 3.0 – Suggested strategy

Balancing these two competing needs – to build expertise and to explore – have been a challenging and valuable activity. In Figure 1 below, the division of exploration and expertise is shown in yellow. To maximize the potential of ML 3.0 to rapidly identify, assess, and roll out ML-related innovations and changes, we suggest the following strategy.

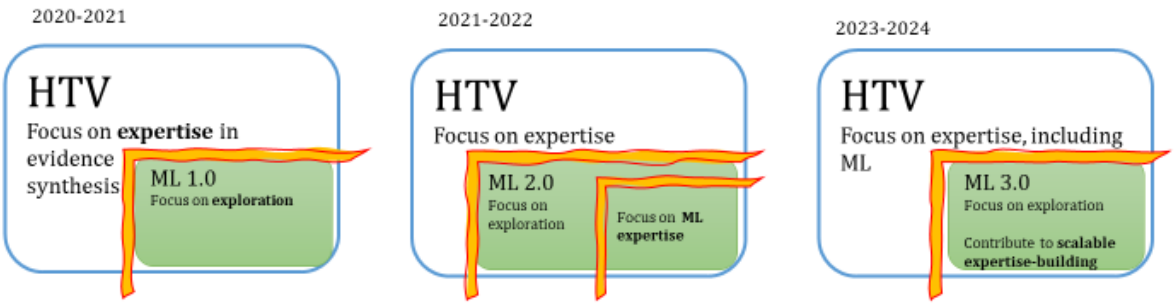

Figure 1: *Division of exploration and expertise in HTV*

ML 3.0 focuses on exploration

HTV should have a system to keep looking outwards to identify new solutions, and there should be support for those looking – that is, some people or groups in HTV must be given the mandate to challenge the status quo of how we produce reports, why, and for whom.

We recommend that ML 3.0 maintain its focus on exploration as one element of HTV's "outward-looking" strategy.

Within the team, exploration activities also require a continuous wish to challenge the status quo of the current suite of ML functions supported (see *Focus areas* below for examples). We must stay as up to date as possible on developments in the field, and we have successful explorations methods to continue with: continuously horizon-scanning the literature, ResearchGate and similar social media sites, staying updated with conference topics and

output, chasing and following up with new topics, and remaining strongly linked to our existing network of experts through formal projects and informal collaboration. Exploration activities all require a high level of independence and knowledge, as a new function or application can only be identified by team members as potentially innovative if they are aware of the existing situation. The pace of innovation activities is fast, and the information flow both "in" to the team and back "out" in the form of recommendations and planned evaluations.

ML 3.0 provides scalable expertise-building for novel functions

ML 3.0 will identify and evaluate new functions and new applications; hence it will remain the team's responsibility to build capacity for these. We suggest that the ML team in general provides training to existing HTV structures (those systems HTV already has in place to teach or upskill employees, see first bullet point list on next page), rather than review teams, and only on new functions.

To a greater extent than before, the capacity-building that ML 3.0 provides must be scalable: it must be repeatable and reach more employees at a time (i.e. cost-effective). Figure 2 displays some example activities from least to most scalable.

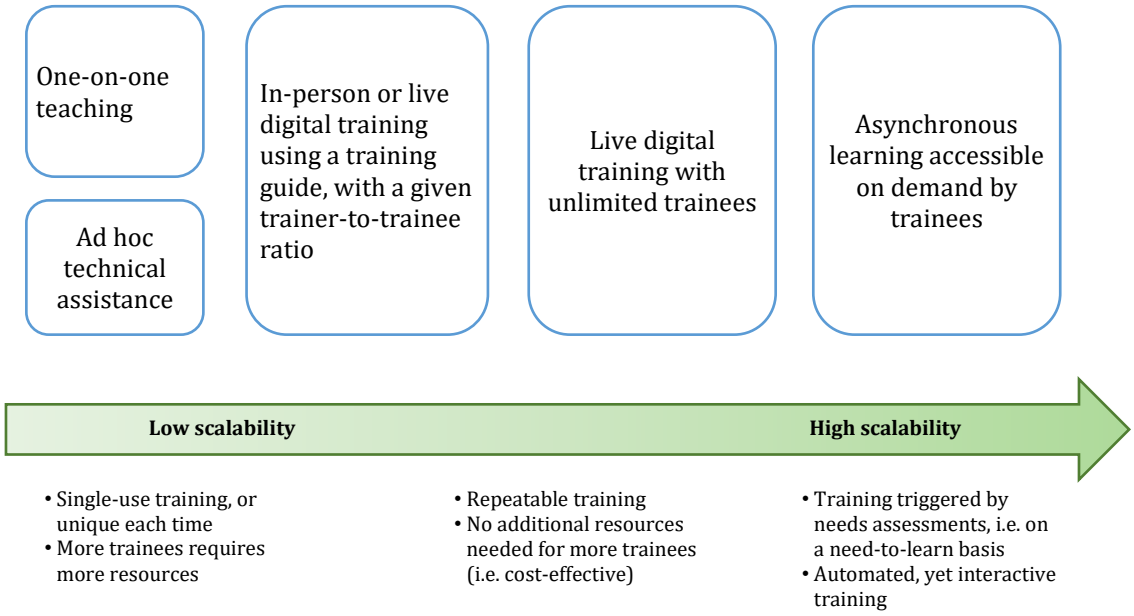

Figure 2: 1 *Examples of capacity-building activities*

HTV adopts responsibility for expertise-building of established functions

For all established functions for which ML 2.0 has identified or created, quality-controlled, and piloted training materials, we recommend that the further use of these training materials be shifted from ML 3.0 to HTV, as displayed in the third panel of Figure 1. ML 3.0 can oversee and assist the hand-off of all new capacity-building materials to HTV, to ensure a smooth procedure with sufficient support and knowledge. Transferring ML skill-building responsibility to established mechanisms (see below) to foster skill-building is a more effective and sustainable use of existing HTV structures. Instead of initiating a separate ML teaching team, building ML skills can be adopted into a variety of existing mechanisms. As the

training material for new functions becomes produced more sustainably, this adoption process will become increasingly easy, as scalable training should not require expert trainers.

Existing mechanisms/structures:

- *New employee training:* Recordings of "ML Week 2022" plus all other new training material could be re-shown or re-given live, by employees who already train new employees.
- *Metodehåndbok*: Descriptive text on ML could be excised from existing training materials and added.
- *Laglеders administrativ håndbok*, checklists for project leads, peer review checklists/guidance, and other quality control standards and documentation: While administrative tools, these standards could also include ML-related steps.
- *Protocol and report templates:* These were first updated in February 2022, and a new round of updates is underway.
- *Undervisningslaget*: This team is already tasked with building capacity to external groups, and as such, is comprised of teaching and methods experts. This team could add ML to their topics, and add HTV employees as a target audience. In addition, there is a clear need in the evidence synthesis field for ML training, and the benefits of the *undervisningslag* hiring themselves out could be substantial: networking with organizations who can contribute to ongoing or planned ML evaluations, the expertise gained by the trainers, and payment or in-kind knowledge exchange.
- *Existing internal networks as arenas for learning*: Recordings of "ML Week 2022" plus all other new training material could be repeated in selected network meetings by network leads or other interested individuals.
- *Acquiring specialists*: Just as NIPH hires statisticians and other methods experts, HTV could aquire ML or AI specialists/expertise from other divisions or from outside NIPH.

ML Team 3.0 as part of a cluster portfolio

Our second suggestion is that HTV think like a futurist: rather than thinking about ML 3.0 as "the" right answer and driving forward with it, ML 3.0 should be seen as one of many scenarios, pathways, and possibilities. To operationalize this, HTV could situate ML 3.0 as only one team in a portfolio of exploration activities.

Characteristics of a cluster portfolio:

- Scope:
  - More exhaustive than ML, and also include explorations around automation, digitalization, and other work flow changes. Any activities that involved exploration rather than expertise-building, and so-called radical change rather than incremental change, could be candidates.
- Content type:
  - Include teams such as ML 3.0 and *metodevarsling*, as well as activities, time-limited inivitatives, concept phase explorations, and so on. By being open to activities that are not organized as teams, the portfolio itself could learn from alternative organizations.

    - The subordinate activities could learn from the infrastructure and lessons learned from ML 2.0 in particular, to the extent useful.
- General skills that may be beneficial for involved employees:
    - Change management
    - Change communication
    - Innovation leadership
    - Scaling up
    - Innovation performance measurement
- Administration:
    - Coordinated by one or two people, one of whom could be a member of KL. The coordination group would need a different mandate than typical team leaders, i.e. higher level roles coordinating the teams/projects within the portfolio, perhaps at a level between lagleder and kontaktpunkt.
    - Key performance indicators and clear success criteria, as well as high tolerance for risk, are recommended.
    - Data flowing in to the coordinator(s) from the portfolio's projects would be used in decision-making about resource distributon and timelines. With a proper management tool, like a dashboard[1], the coordinator(s) could easily see ongoing risk assessments, status reports, and areas of overlap that could be better exploited.
- Anticipated benefits:
    - Identifying the included activities (see Figure 3 for example portfolio activities) as *exploration* activities would give clear permission for the different expectations and conditions that they need to succeed.
    - If the portfolio's content is connected (if each project in the portfolio had access or continuous overview of all projects included), HTV could gain an arena for dedicated learning and support – a supportive sandbox, so to speak. The coordinator(s) as well as activity leads, if not all or the majority involved, could meet regularly for the express purpose of learning from each other's challenges related to change management and innovation adoption.
    - Identify and grow employees and their change and innovation skills.
    - Incubator for funding applications.

HTV is well-situated to create such a portfolio. It already supports a number of development initiatives, and has a deep pool of creative and skilled employees from which to draw. Figure 3 below displays only some examples of activities that could be seen as part of this portfolio.

[1] A software where a portfolio coordinator can log on and see the status of multiple projects at once – progress, late tasks, hours used, overall statuses, etc.

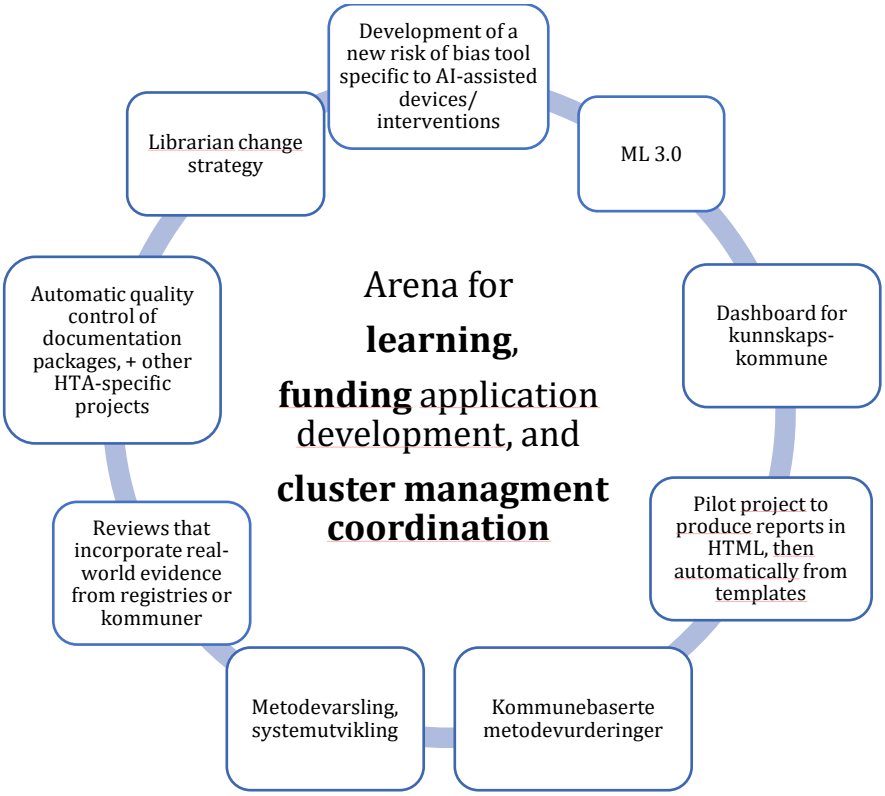

Figure 3: *Example activities in an exploration portfolio*

Focus areas

The focus areas below apply to ML Team 3.0 but could also be included as distinct activities in a cluster-wide portfolio.

Products: Review updates and living reviews

What we know

-Evidence-informed decision making transcends summarizing the evidence. It also implies using scare resources more efficiently, accountability of decisions, and reducing research waste.

-Living reviews - a term that can be applied to any types of evidence synthesis - are highlighted by WHO as "a methodology that can help improve timeliness and quality as they use systematic review quality methods but are frequently updated to ensure that they are also current" (2).

-Living reviews help us better respond to commissioners' and other stakeholders' needs without underminig quality. Their workload savings may be used to capture evidence on contextual factors, such as equity or acceptability. This will ultimately enhance the usefulness of our reports.

-ML (used mainly for study selection) likely has an important role in making living reviews sustainable. It can be a way to compensate for the enormous workload of updating reviews (3) .

-HTV has gained fruitful experience in the production of living reviews in different areas, such as COVID-19 (Omicron living map) and social welfare.

Potential

-Semi- to full-automation of living review products, e.g. (4).

-Flexible. Could be used to accommodate changes in PICO made by commissioner.

-Living reviews may enhance commissioners' satisfaction given their potential to permit the study of contextual factors via a more efficient use of resources.

-HTV will be able to adapt and align to the fast-pace dynamics of evidence synthesis worldwide.

-Increased motivation among reviewers at HTV, as they will explore new methods to make meaningful changes happen.

Suggested next steps

-Strengthenen the use of OpenAlex and custom classifiers to regularly update and populate living evidence maps as well as reviews.

-Aim for a close dialogue with both commissioners and end users about the potential benefits and uses of living reviews to meet new and existing needs.

-Explore (in our context) the already documented relevance of living reviews in facilitating contextualized decision-making scenarios, similar to ongoing work in the *kunnsksapskommune*

-Join efforts with other institutions towards living repositories of evidence (PICOs or Evidence profiles needed for EtD frameworks). This is similar to what Epistemonikos has done with L.OVE.

-Potentially re-visit the data-sharing initiative. This was de-prioritized by the *Digitaliseringsportfølje* in 2021, but all project management material and funding applications can be re-used.

Process: Novel review presentation or product forms

What we know

-Reports continue to be written in Microsoft Word based on templates for specific products from the product portfolio. There is no automation and a high degree of formatting needed from the user. Navigating the process of delivering a product is complex and requires a good overview of methodological approaches as well as the institutional guidelines, and integration of new methods as ML is not intuitively linked to specific products.

-Together these aspects form a high barrier with a steep learning curve for new employees, as well as provide challenges to keeping everyone up to date on current guidelines and practises.

-A more automated approach, based on a decision tree, could guide new and current employees through the process of writing a report, take care of the formating and secure that guidenlines, checklist and methods are appropriately used.

Potential

-Such an approach could create HTML reports, and save data in a more accessible format reducing duplication as well as improving readibility.

-New products, especially simpler products can be developed and rolled out directly.

-A more user-friendly approach, requring less oversight to get started, can be a low threshold entrance to new employees and new players to the field as for example the low resource and capacity setting found in the kommune-level. metodevurderinger

-An underlying structure behind all products would make it easier to index reports.

Suggested next steps

Discussion of how workflow processes can be adjusted to the production line, and an evaluation of the technical feasibility, barriers, and risks associated with such an approach.

Specific function: ChatGPT and advanced natural language processing

What we know

ChatGPT is a chatbot powered by an advanced language model, GPT-3.5. While not a knowledge model, a large amount of knowledge has been used in order to train it. Users can engage with ChatGPT and request feedback, answers, and information as if they were conversing with a human, but ChatGPT is not intended to provide "perfect" information. It was trained from information harvested from the internet primarily before 2021.

We have used ChatGPT to draft the Key Messages of this report automatically, by copying the text of the Suggested strategy section and requesting a summarized version. We then requested a Norwegian translation of the summary, which became our *Hovedbudskap*.

ChatGPT is free and available online for research preview, as of Dec. 2022. Given the global demand, it is not unlikely that a fee-for-service model will be introduced.

Potential

As a language model, ChatGPT (and others) offer us the ability to automatically produce text that is semantically meaningful and undifferentiable from text produced first by an employee. Some immediate applications:

- Draft *Omtaler* directly from the text of a selection of literature.
- Draft Key messages, Executive Summaries, and other summaries of our reports.
- Translate our text.
- Simplify texts for different target audiences
- Conduct risk of bias or methodological quality assessments as an initial step or as an independent reviewer.

These applications are quite conservative.

An overall potential of ChatGPT is to significantly reduce the human effort needed to produce high-quality text. The produced text can be a summary, an analysis, a translation, or something else.

Suggested next steps

-Explore the applications brainstormed above.

-Conduct a rapid study comparing ChatGPT-produced Cochrane Risk of Bias assessments against human-generated assessments.

-Create a communication and education strategy.

-Assess resource use and potentially reserve resources for otaining access in the future.

Review phase: Data extraction

What we know

We have managed to greatly improve efficiency of the study selection process, due to incorporation of ML functions in our workflow. A natural next step would be to explore ML alternatives for data extraction. ML can be used to identify and extract information (semi)automatically, which has the potential to streamline this process. Information extracted can be study characteristics (typically your table 1 descriptive content) and study findings.

Potential

Data extraction is often time-consuming, resource-intensive and repetitive, due to it currently being a largely manual process as well as the complexity and amount of data needed to be included in our reviews. Hence it is an ideal candidate for ML use, both before and after screening.

Before screening: For many of our systematic reviews, we are only interested in publications from certain countries, e.g. OECD countries or the Nordic countries, and therefore it would be very helpful to find a relatively quick and systematic way to easily identify the origin country for each reference before starting to screen the references. Geoparsing can be used for this type of process, which involves converting free-text descriptions of places (e.g. provided in title/abstracts) into geographic identifiers. Software packages for geoparsing is available in java and in python. However, these are unlikely standalone candidates for large scale implementation at NIPH, as use require programming skills.

After screening: ML alternatives for data extraction are much scarcer than ML for screening, particularly for non-clinical reviews. However, there are at least two web-based applications that incorporates ML based semi-automated data extraction: Colandr and Dextr.

Suggested next steps

-Further explore the possibilities of using Colandr and Dextr.

-Maintain collaboration with JKI, the evidence synthesis organization and developers of the free review software Cadima. They are currently currently assessing resource requirements for developing a tool allowing for direct text annotation and extraction. The ML team has provided input on user requirements. See the ML 2.0 final report for more details.

-Develop in-house a tool with easy to use GUI for geoparsing, by creating a Shiny app. Creating a shiny app is free, but we would need to find to expertise to create it, such as a programmer or existing employee with these skills.

Conclusion

We fully support HTV's continued investment in ML, as part of its strategic innovation.

When HTV reviews are "saturated" with the most recent ML functions, and employees are able to identify, understand, and critically use up-and-coming, novel functions themselves, the ML team can completely transition into another innovation team. Even when ML itself is no longer a focus, the need for an iterative team as a driver of change will remain. The infrastructure of ML 3.0 may be readily used as a change agent for future advances.

HTV can champion ML 3.0 as one of many innovation activities in a larger portfolio to build best practices and organizational options to integrate innovations most effectively.

References

1. De Wit B. Strategy : an international perspective. Seventh edition. ed. Andover: Cengage; 2020.
2. World Health Organization. Evidence, policy, impact: WHO guide for evidence-informed decision-making. Geneva: World Health Organization; 2021. Available from: https://apps.who.int/iris/handle/10665/350994
3. Tercero-Hidalgo JR, Khan KS, Bueno-Cavanillas A, Fernández-López R, Huete JF, Amezcua-Prieto C, et al. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. Journal of Clinical Epidemiology 2022;148:124-34.
4. Shemilt I, Noel-Storr A, Thomas J, Featherstone R, Mavergames C. Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier. Systematic Reviews 2022;11(1):15.

Published by the Norwegian Institute of Public Health
February 2023
P.O.B 4404 Nydalen
NO-0403 Oslo
Phone: + 47-21 07 70 00
The report can be downloaded as pdf
at www.fhi.no/en/publ/