SCOR: A secure international informatics infrastructure to investigate COVID-19

JL Raisaro[1]; Francesco Marino[2]; Juan Troncoso-Pastoriza[2]; Raphaelle Beau-Lejdstrom[4]; Riccardo Bellazzi[5,20]; Robert Murphy[3], Elmer V. Bernstam[3, 23]; Henry Wang[24], Mauro Bucalo[21]; Yong Chen[6]; Assaf Gottlieb[3]; Arif Harmanci[3]; Miran Kim[3]; Yejin Kim[3]; Jeffrey Klann[7]; Catherine Klersy[8]; Bradley A. Malin[9]; Marie Méan[10]; Fabian Prasser[11,12] ; Luigia Scudeller[13]; Ali Torkamani[14]; Julien Vaucher[10]; Mamta Puppala[15]; Stephen T.C. Wong[15]; Milana Frenkel-Morgenstern[16]; Hua Xu[3]; Baba Maiyaki Musa[22]; Abdulrazaq G. Habib[22]; Trevor Cohen[17], Adam Wilcox[17]; Hamisu M. Salihu[18]; Heidi Sofia[19]; Xiaoqian Jiang[3]; JP Hubaux[2].

[1]Data Science Group and Precision Medicine Unit, Lausanne University Hospital, Lausanne, Switzerland

[2]Laboratory for Data Security, EPFL, Lausanne, Switzerland

[3]School of Biomedical Informatics, UTHealth, Houston, Texas

[4]Institute of Global Health, University of Geneva, Geneva, Switzerland

[5]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

[6]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania

[7]Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts

[8]Biometry and Clinical Epidemiology Service, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

[9]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee

[10]Department of Internal Medicine, Lausanne University Hospital, Lausanne, Switzerland

[11]Medical Informatics Group, Berlin Institute of Health, Berlin, Germany

[12]Charité - Universitätsmedizin Berlin, Berlin, Germany

[13]Scientific Direction, Clinical Epidemiology and Biostatistics, Fondazione IRCCS Ca' Grande Ospedale Maggiore Policlinico, Milan, Italy

[14]Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA USA

[15]Department of Systems Medicine and Bioengineering, Houston Methodist Cancer Center, Weill Cornell Medical College, Houston, Texas

[16] Cancer Genomics and BioComputing of Complex Diseases laboratory, Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel

[17] Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington

[18] Department of Family and Community Medicine, Baylor College of Medicine, Houston, Texas

[19] National Institutes of Health (NIH) - National Human Genome Research Institute, USA

[20] IRCCS ICS Maugeri, Pavia, Italy

[21] BIOMERIS srl, Pavia, Italy

[22] Department of Medicine, Africa Center of Excellence in Population Health and Policy, Bayero University Kano, Nigeria

[23] Division of General Internal Medicine, Department of Internal Medicine, McGovern School of Medicine, UTHealth, Houston, Texas

[24]Department of Emergency Medicine, McGovern School of Medicine, UTHealth, Houston, Texas

Abstract:
Global pandemics call for large and diverse healthcare data to study various risk factors, treatment options, and disease progression patterns. Despite the enormous efforts of many large data consortium initiatives, the scientific community still lacks a secure and privacy-preserving infrastructure to support auditable data sharing and facilitate automated and legally compliant federated analysis on an international scale. Existing health informatics systems do not incorporate the latest progress in modern security and federated machine learning algorithms, which are poised to offer solutions. An international group of passionate researchers came together with a joint mission to solve the problem with our finest models and tools. The SCOR consortium has developed a ready-to-deploy secure infrastructure using world-class privacy and security technologies to reconcile the privacy/utility conflicts. We hope our effort will make a change and accelerate research in future pandemics with broad and diverse samples on an international scale.

Mission

A major lesson that the COVID-19 pandemic has already taught the scientific community is that timely international data sharing and collaborative data analysis is absolutely vital to navigate through policy decisions that have life-or-death consequences. Some of the most pressing issues about COVID-19 infections require urgent sharing of high-quality data concerning, for example, risk factors that influence infection, prognosis, and predictions of drug response from phenotypic, genotypic, and epigenetic data [1]. To generate or test scientific hypotheses, we need large-scale and well-characterized patient-level datasets to provide sufficient statistical power. Building and sharing massive datasets containing personal health information have numerous legal and ethical implications that hinder new discoveries and prevent the scientific community from assessing their validity [2]. In this respect, the case of two COVID-19 related articles published by The Lancet [3] and The New England Journal of Medicine [4] serves as an example. When concerns were raised regarding the veracity of the data used to support the conclusions in these articles, the two prestigious journals requested access to the raw data to conduct independent reviews. However, the authors could not comply with such a request, as granting access to the data would have violated confidentiality requirements, and the two journals had no choice but to retract the articles [5,6]. These instances reinforce the need for a robust privacy- and confidentiality-compliant data processing and sharing system to address these challenges in the era of COVID-19 and future pandemics.

Numerous data-driven projects have been launched across the globe to combat COVID-19, as summarized below. Yet, there is a lack of systematic support to address one of the main impediments that prevent and delay broad and sustainable medical data sharing: privacy protection. To address privacy protection challenges, researchers make trade-offs on data utility. On the one hand, several data-sharing projects on COVID-19 are based on a decentralized approach, employing the computation of local statistics (sometimes obfuscated to hide small numbers) that are subsequently shared and aggregated through meta-analysis.

However, case numbers may sometimes be too low in certain subpopulations and could be considered identifiable information, which can make it very challenging for hospitals to even share aggregated data. Additionally, the approach only offers limited results and often depends on voluntary local analyses with human-in-the-loop approval and execution. On the other hand, other projects aim to centralize patient-level data from COVID-19 at a single site and then perform the analysis. Yet, that approach does not easily scale to international collaborations due to the heterogeneity and potential incompatibility of the various legal frameworks. We believe that there are more effective and privacy-congruent solutions to deal with this long-standing challenge and that privacy-by-design technology should be developed and is recently available for deployment to address the utmost urgency of data sharing by reducing administrative and regulatory barriers driven by privacy and security concerns. With this goal in mind, we have established an international consortium for Secure COllective Research (SCOR) [7] to deploy the next-generation distributed infrastructure and tools for secure data sharing, analysis, and mining while respecting patient privacy and maximizing data utility during global disease outbreaks like the current COVID-19 pandemic. The list of founding partners for this global initiative is provided in Supplement S1.

Short- and long-term goals

SCOR aims to achieve the following goals:

- Short-term: establish a proof-of-concept decentralized and privacy-preserving analytics platform, taking advantage of world-class privacy technology for COVID-19 data supporting cohort exploration for assessing the feasibility of research study protocols, and facilitating speedy patient cohort recruitment.

- Long-term: build a distributed privacy-preserving and sustainable infrastructure for federated statistical and machine learning analysis to support multi-center clinical studies of the COVID-19 outbreak and future pandemics.

Positioning of SCOR regarding other similar initiatives

SCOR is a new initiative that complements existing multi-centric data-sharing efforts to face the COVID-19 pandemic. COVID-19 research moves rapidly with new initiatives announced daily. In Table 1 we summarize the major initiatives we are aware of (as of June 2020) and compare them to SCOR along the following axes:

- Type of analyses (cohort exploration vs. meta-analysis vs. distributed analytics vs. centralized analytics)
- Data storage (centralized vs. decentralized)
- Scope (national vs. international)
- Type of data transferred (aggregate-level vs. patient-level)
- Data protection mechanism (local obfuscation, global obfuscation, encryption)
- Level of automation (manual analysis, semi-automated analysis, fully automated system)

The approach proposed by SCOR is the only one that (i) provides operational continuity for the long run, as it relies on a fully automated software platform for distributed data sharing, (ii) has an international scope, and (iii) provides the best data privacy/utility trade-offs, as it enables both cohort exploration and distributed analytics under strong privacy guarantees. These guarantees are ensured by deploying encryption techniques for distributed secure information aggregation across sites, lowering the need for local obfuscation.

Clinical research goals

The rapid spread of the COVID-19 epidemic globally has almost overwhelmed health systems worldwide and it has already claimed lives in the hundreds of thousands. Starting from Asia, followed by Europe and next by the rest of the world, the first wave is now decreasing. No treatment has yet been demonstrated to be unequivocally effective and the subpopulation stratification of disease risks is still lacking, with multiple facets of presentation and prognosis. In particular, the recognized initial respiratory signs, symptoms, and laboratory findings have extended to many other settings, including dermatology, neurology, and hematology. Hospitals around the world have set up COVID-19 registries to accumulate information on symptoms, laboratory, respiratory function, imaging, and treatment to understand the disease. Joining forces will increase the number of patients that can be analyzed to address the next wave of the pandemic. Data harmonization will be challenging, but ultimately essential. Similarly, the proposed secure and distributed data analysis approach will overcome obstacles to information sharing which some institutions are often reluctant to do. The SCOR network will serve as a hub for bringing together clinical research groups based on shared interests.

To demonstrate the utility of the SCOR approach, we will develop and apply use case scenarios (Box 1) that require data aggregation across multiple sites as each site has only a narrow view of the required information. This partial view stems from the uniqueness of the population at each site and from the difference in research protocols applied at each site.

SCOR requirements and existing data-sharing platforms

The aim of SCOR is to provide an ecosystem for privacy-preserving distributed data analysis, which addresses all the five dimensions of secure data management, as expressed in the Five Safes framework [15] (safe projects, safe people, safe setting, safe data, safe outputs) while overcoming the loss of data utility typical of existing decentralized approaches based on study-level meta-analyses that rely on site-level (i.e., local) obfuscation to protect patients' privacy. We distinguish between (i) safes that must be addressed at the consortium level, i.e., safes that are enacted by decisions taken by the SCOR board (representative members from each participating institution) to pursue the high-level consortium's privacy and security goals, and (ii) safes that must be addressed at the platform level, i.e., safes that are enacted by technical safeguards featured by the technological infrastructure of the SCOR analysis platform. More details about the rational and platform requirements are discussed in Supplementary S0.

Table 2 briefly summarizes the most wide-spread distributed medical data analytics platforms in terms of provided functionalities and protection mechanisms to ensure safe settings and safe output requirements. We focus our comparison on the public platforms as they allow for an in-depth analysis. Yet, there exist also proprietary/closed platforms such as TriNetX, InSite, and Clinerion that, to the best of our knowledge, only partially address the data protection requirements for the SCOR initiative.

Proposed platform: MedCo

Given the SCOR platform requirements, the MedCo analysis platform [16] is the one that best addresses them (Figure 1).

Privacy-preserving technological enablers

Homomorphic encryption

Homomorphic encryption (HE) [17] supports computation on encrypted data (ciphertexts). Thanks to this property, homomorphically encrypted data can be safely handed out to third parties, who can perform meaningful operations on them without learning anything about their content. While fully homomorphic encryption schemes, i.e., schemes that enable arbitrary computations on ciphertexts, are still considered non-viable due to the high computational and storage overheads they introduce, practical schemes that enable only a limited number of computations on ciphertexts, e.g., additions and multiplications, have reached a level of maturity that enables their use in real scenarios.

Secure multi-party computation

Secure multi-party computation (SMC) [18] protocols allow multiple parties to jointly compute functions over their private inputs (e.g., confidential patient-level data) without disclosing to the other parties more information about their inputs than what can be inferred from the output of the computation. This class of protocols is particularly attractive in privacy-preserving distributed analytic platforms due to the great variety of secure computations they enable. However, this flexibility often comes with a number of drawbacks that hinder their adoption, including high network overhead, and parties required to be online during the computation. HE and SMC can be fruitfully employed in combination to mitigate their respective overheads and limitations and to provide effective solutions for privacy-preserving distributed analysis on sensitive data.

Data obfuscation

Data obfuscation techniques reduce the input data detail to an acceptable minimum and limit the information leakage stemming from the disclosure of the results. Indeed, even if data are kept private, the results of analyses performed may still reveal information about subjects that can be used to infer sensitive properties. Data obfuscation techniques alter data in a deterministic manner (e.g., k-anonymity [19], often applied to input data) or statistical manner (e.g., differential privacy [20], often implemented into processing methods to ensure safe outputs). For the results to remain useful, the amount of noise introduced by data obfuscation has to be carefully calibrated to reach the desired trade-off between utility and privacy. Studies show that k-anonymity and differential privacy sometimes give disappointing results when the target sample size is small [21,22]. It is not a problem of both mechanisms but the unavoidable challenges in maneuvering statistics with limited flexibility. This issue is alleviated when safe settings are used to create large (protected) virtual datasets, compared to applying data obfuscation to local datasets.

Operating principles

By using MedCo, health professionals and scientists can query data scattered among diverse institutions as if it were stored in one single location (virtual collective dataset), but without the need of seeing nor moving the data. As such, it facilitates compliance with stringent data protection regulations such as the EU GDPR [23] and the US HIPAA [24]. We include details about access control and accountability in supplement S7 and SCOR deployment plan in *Supplement S8*.

Ethical issues

Ethical issues in data sharing and analysis are on the rise. Our technology provides privacy and security safeguards to automate global information exchange, but it might make the direct assessment of healthcare disparity harder due to the obfuscation. Fairness, equity, and transparency of medical informatics models represent the fundamental considerations for public trust and clinical usability. Many seemingly objective models are indeed influenced by their design, which can significantly over- or under-estimate the risks on different subpopulations and introducing an unjustified basis for discriminating against a subpopulation. Such problems might be aggravated in a federated network with strong security protection, if unnoticed, could result in significant ethical challenges. As a community, we should take a high standard in addressing these problems by-design to consider fairness, equality, and justice to conduct responsible medical research.

Conclusion

There are urgent needs for data sharing and analysis in COVID-19, but we should not give up privacy in responsible research under pandemics. It is crucial to work together and build a robust and scalable infrastructure with state-of-the-art security and privacy technology to enable automated federated data analysis to accelerate scientific discoveries to combat the SARS-CoV-2 outbreak and future pandemics. We are fully committed to establishing this international consortium of collective data and knowledge discovery network to support clinical research to answer important questions.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. Co-author Milana Frenkel-Morgenstern is supported by COVID-19 Data Science Institute (DSI), Bar-Ilan University (grant number 247017).

Competing Interests Statement:

Riccardo Bellazzi is a shareholder of Biomeris s.rl. Hua Xu have financial related interest at Melax Technologies Inc. Raphaelle Beau-Lejdstrom serves as a Real World Evidence consultant for Pharmaceutical industry (UCB Pharma). The other co-authors have no competing interests to declare.

Contributorship Statement:

Raisaro, Hubaux, Malin, Gottlieb, and Jiang were responsible for the conception and design of the paper. Raisaro, Hubaux, Jiang, Bernstam, Frenkel-Morgenstern, Gottlieb, Troncoso-Pastoriza, Marino, Chen, Malin, Klann, and Sofia drafted the paper. Murphy, Puppala, Wong, Bernstam, Frenkel-Morgenstern, Musa, Habib, Wilcox, Bucalo, Gottlieb, Torkamani, Méan, Vaucher, Klersy, Scudeller, and Salihu acquired and contributed data, participated in the discussion, and edited/reviewed the manuscript. Klann, Malin, Xu, Marino, Troncoso-Pastoriza, M. Kim, Chen, Sofia, Prasser participated in the idea discussion and reviewed/edited/contributed to the manuscript. Bellazzi, Beau-Lejdstrom, Harmanci, Y. Kim, Wang, Cohen reviewed the manuscript and conducted the final approval of the version to be published. All authors agreed to submit the report for publication.

References

1 How sick will the coronavirus make you? The answer may be in your genes. Science | AAAS. 2020. doi:10.1126/science.abb9192

2 Sittig DF, Singh H. COVID-19 and the Need for a National Health Information Technology Infrastructure. *JAMA* Published Online First: 18 May 2020. doi:10.1001/jama.2020.7239

3 Mehra MR, Desai SS, Ruschitzka F, *et al.* Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* Published Online First: May 2020. doi:10.1016/S0140-6736(20)31180-6

4 Mehra MR, Desai SS, Kuy S, *et al.* Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. *N Engl J Med* Published Online First: 1 May 2020. doi:10.1056/NEJMoa2007621

5 Mehra MR, Ruschitzka F, Patel AN. Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* Published Online First: 5 June 2020. doi:10.1016/S0140-6736(20)31324-6

6 Mehra MR, Desai SS, Kuy S, *et al.* Retraction: Cardiovascular Disease, Drug Therapy, and Mortality

in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. *N Engl J Med* Published Online First: 4 June 2020. doi:10.1056/NEJMc2021225

7 Secure Covid Research | Secure Collective Covid-19 Research. https://securecovidresearch.org/ (accessed 5 May 2020).

8 COVID19 | Cancer Genomics and BioComputing of Complex Diseases Lab. Cancer Genomics and BioComputing Lab. 2020.http://mfm-lab.md.biu.ac.il/research/covid19 (accessed 5 May 2020).

9 Funk MJ, Westreich D, Wiesen C, *et al.* Doubly robust estimation of causal effects. *Am J Epidemiol* 2011;**173**:761–7.

10 Li L, Huang T, Wang Y, *et al.* COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. Journal of Medical Virology. 2020;**92**:577–83. doi:10.1002/jmv.25757

11 Richardson S, Hirsch JS, Narasimhan M, *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* Published Online First: 22 April 2020. doi:10.1001/jama.2020.6775

12 Cui S, Chen S, Li X, *et al.* Prevalence of venous thromboembolism in patients with severe novel coronavirus pneumonia. *J Thromb Haemost* Published Online First: 9 April 2020. doi:10.1111/jth.14830

13 Klok FA, Kruip MJHA, van der Meer NJM, *et al.* Incidence of thrombotic complications in critically ill ICU patients with COVID-19. *Thromb Res* Published Online First: 10 April 2020. doi:10.1016/j.thromres.2020.04.013

14 Helms J, Tacquard C, Severac F, *et al.* High risk of thrombosis in patients with severe SARS-CoV-2 infection: a multicenter prospective cohort study. *Intensive Care Med* Published Online First: 4 May 2020. doi:10.1007/s00134-020-06062-x

15 Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. Published Online First: 2016.https://uwe-repository.worktribe.com/output/914745

16 MedCo | Collective protection of medical data. https://medco.epfl.ch/ (accessed 13 Apr 2020).

17 Gentry C. Fully homomorphic encryption using ideal lattices. In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009. 169–78.

18 Shaikh Z, Garg P. Secure Multiparty Computing Protocol. Interdisciplinary Perspectives on Business Convergence, Computing, and Legality. 2013;:132–43. doi:10.4018/978-1-4666-4209-6.ch012

19 Sweeney L. k-anonymity: A model for protecting privacy. *International journal on uncertainty, fuzziness and knowledge-based systems* 2002;**10**:557–70.

20 Dwork C. Differential privacy. In: *Encyclopedia of Cryptography and Security*. Springer 2011. 338–40.

21 Vaidya J, Shafiq B, Jiang X, *et al.* Identifying inference attacks against healthcare data repositories. *AMIA Jt Summits Transl Sci Proc* 2013;**2013**:262–6.

22 Bambauer J, Muralidhar K, Sarathy R. Fool's gold: an illustrated critique of differential privacy. *Vand J Ent & Tech L* 2013;**16**:701.

23 General Data Protection Regulation (GDPR) Compliance Guidelines. GDPR.eu. https://gdpr.eu/ (accessed 5 May 2020).

24 Health Insurance Portability and Accountability Act (HIPAA). http://www.hhs.gov/ocr/hipaa

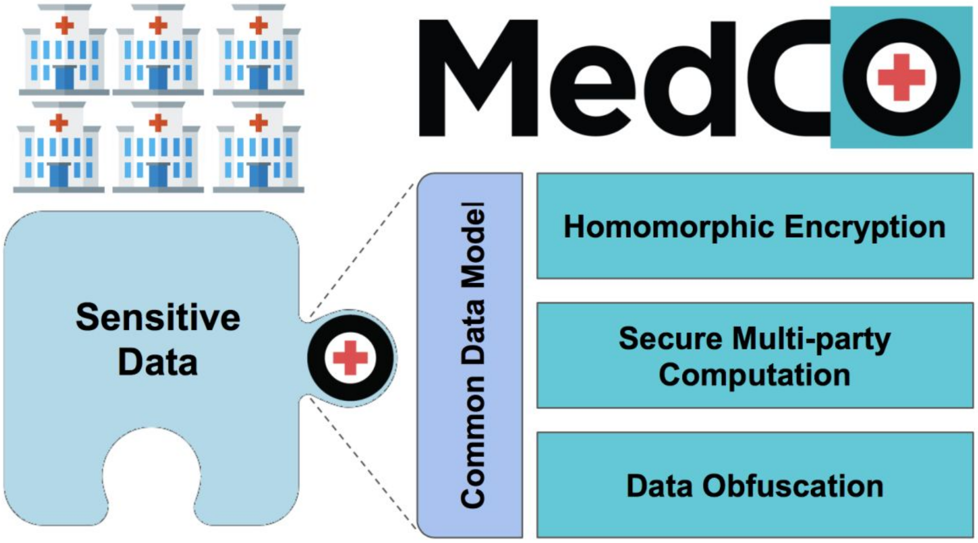

Figure 1. MedCo core technologies. MedCo is a decentralized software system that uses cutting-edge privacy-preserving technologies to enable the secure sharing of medical data among health institutions. It builds on three core privacy-preserving technologies: homomorphic encryption, secure multi-party computation, and data obfuscation. These technologies are used in synergy to combine information owned by multiple institutions and reveal otherwise hidden global insights while addressing legal and privacy concerns.

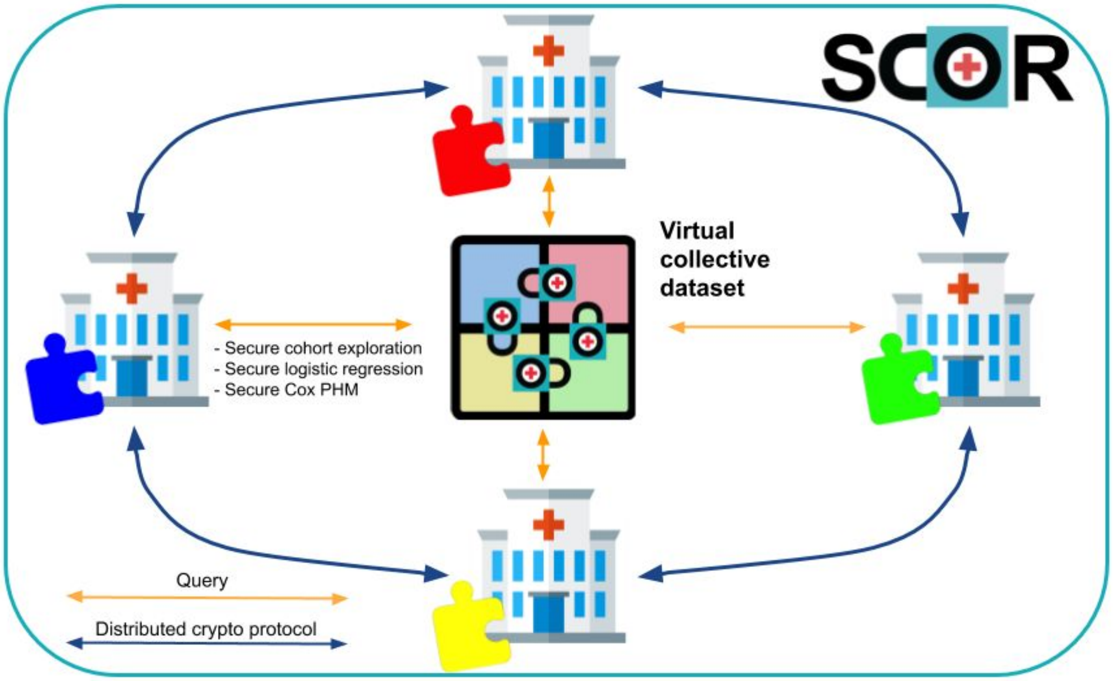
SCOR
Virtual
collective
dataset
- Secure cohort exploration
- Secure logistic regression
- Secure Cox PHM
Query
Distributed crypto protocol

Figure 2. The SCOR MedCo approach: when an institution queries the virtual collective dataset, it engages in a distributed cryptographic protocol with all the other institutions to securely obtain the result of the query. MedCo provides end-to-end protection against unauthorized access to data thanks to homomorphic encryption, which allows keeping the data in an encrypted state not only at rest and in transit but also during computation (safe settings). MedCo also removes the need for a central trusted authority by leveraging secure multi-party computation. The result of a query/analysis can be decrypted only through a distributed protocol that involves the approval of all the participating institutions. If one or more institutions are compromised by a cyberattack, the others can refuse to decrypt the data, thus keeping the data secure.

Box 1: Demonstrative research study protocols that are planned to be conducted on the SCOR secure infrastructure

Use case 1: Risk stratification for COVID-19 patients

We will collect patient demographics (Sex, Age, Race/Ethnicity), smoking status, vitals and/or their fluctuation over time (BMI, oxygen saturation, and blood pressure), comorbidities (Diabetes, Lung Disease, Cancer, Immunodeficiency, Heart Disease, Hypertension, Asthma, Kidney Disease, and GI/Liver Disease) and the outcome (length of stay in hospital or in ICU, discharge or death) and apply multivariate (non-linear) machine learning classifiers to create a personal risk score that accounts for regional differences.

Use case 2: Efficient treatments for COVID-19 patients

We will collect candidate medications assembled manually curated by the Bar-Ilan University in Israel from trials and studies [8] and study their effectiveness in treating COVID-19 patients. Using doubly robust methods that integrate standardization and inverse probability weighting techniques [9] (considering time-dependent treatments, left-truncation, interventions like ventilators and Extracorporeal Membrane Oxygenation (ECMO), demographics, smoking status, and comorbidities), we will study averaged treatment effects on the treated (ATT) and conduct time-to-event analysis on mortality, respiratory failure, ICU admission, and length of hospitalization.

Use case 3: Hospital readmission risk factors and prediction of post-hospitalization COVID-19 patients

Despite COVID-19 can cause severe respiratory failure and death, the majority of patients hospitalized for COVID-19 are discharged alive, amounting to 50% in China and 80% in the US [10] [11]. Whether COVID-19 discharged patients are at increased risk of hospital readmission remains unknown as there is no available data regarding the readmission rate of COVID-19 inpatients at 30 days yet. Similarly, the impact of COVID-19 pandemic on hospital readmission of non-COVID-19 patients is unknown. We aim at assessing readmission risk during coronavirus outbreak in medically hospitalized patients and whether COVID-19 inpatients are at increased risk of readmission compared to non-COVID-19 inpatients. This information can be used as a proxy for the quality of health care systems and will provide crucial information on the capacity of different health systems to respond to a global sanitary problem, whether linked to a subsequent wave of COVID-19 infection or any future pandemic.

Use case 4: Changes in the characteristics of COVID-19 over time

It is a common observation in the western hospitals that COVID-19 patients are not the same in May as they were at the beginning of the pandemic in March. The severity of the hospitalized patients is decreasing, while some complications, such as venous thromboembolism [12–14], might be increasing due to increased medical awareness. Making use of claims data first and registry data next, we may be able to use a multivariate and machine learning approach to model this particular phenomenon with many implications for health organizations and decision-makers.

Use case 5: Host genetics in previously healthy COVID-19 life-threatening patients

The clinical presentation of COVID-19 ranges from mild respiratory symptoms to severe progressive pneumonia, multiorgan failure, and death. A variety of risk factors have been associated with severe COVID-19, but extremely severe clinical presentations of COVID-19 are also observed in young patients with no comorbidity. The identification and characterization of rare genetic variants responsible for the most severe forms of SARS-CoV-2 infection in otherwise healthy individuals will help uncover the genes and pathways that play a crucial role in viral pathogenesis and in antiviral response, which will inform drug and vaccine development.

Table 1. Comparison of SCOR with similar data-sharing initiatives. (*) Comparison of fully automated systems for COVID-19 data sharing is reported in Table 2 below.

Initiative	Type of analysis	Data storage	Scope	Type of data transferred	Data protection mechanism	Level of automation
4CE	meta-analysis	decentralized	international	aggregate-level	local obfuscation	manual analysis
ACT Network	cohort exploration	decentralized	national (USA)	aggregate-level	local obfuscation	fully automated system (SHRINE*)
LEOSS	centralized analytics	centralized	international (only EU)	patient-level	anonymization	manual analysis
OHDSI	meta-analysis	decentralized	international	aggregate-level	local obfuscation	manual analysis
PCORNet CDRNs	meta-analysis	decentralized	national (USA)	aggregate-level	local obfuscation	manual analysis
N3C	centralized analytics	centralized	national (USA)	patient-level	anonymization	manual analysis
SCOR	cohort exploration & decentralized analytics	decentralized	international	aggregate-level	encryption & global obfuscation	fully automated system (MedCO*)

Table 2. Comparison between available medical distributed analysis platforms.

	Functionalities		Safe settings	Safe output	
Platform	**Cohort exploration**	**Distributed analytics**	**Secure Aggregation**	**Local obfuscation**	**Global obfuscation**
SHRINE	●			●	
Medical Informatics Platform	●	●			
DataShield	●	●		●	
MedCo	●	●	●	●	●