Follow-up report: Analisi e raccomandazioni per la fase di riapertura

a cura del SGdL 7: “Big Data & AI for policy” della Task force data-driven anti Covid-19 del MID -- 14/05/2020

Il sottogruppo di lavoro 7 “**Big Data & AI for policy**” della task force data-driven del MID ha lavorato dalla fine di Marzo 2020 con l’obiettivo di proporre metodi e strumenti per la definizione e l’attuazione di politiche basate sui dati e sull’evidenza informativa, sfruttando tecnologie innovative di big data analytics e intelligenza artificiale. Le indicazioni del sottogruppo 7, contenute nel report del 14/04/2020, comprendono **raccomandazioni per affrontare le prossime fasi dell’epidemia** e identificano le principali **esigenze informative a supporto delle decisioni epidemiologiche, cliniche e più in generale di gestione della crisi**. Le indicazioni sono focalizzate su 4 linee di intervento in cui le tecnologie di Big Data & AI, in alcuni casi già in uso, in altri casi da incentivare, possono fornire un contributo. Le linee di intervento prevedono, sia nel loro disegno che nella loro operazionalizzazione, la partecipazione di diversi stakeholders istituzionali a livello regionale e nazionale.

---

1. **Sorveglianza epidemiologica sul territorio**: potenziare con personale e tecnologie i presidi sanitari sul territorio (servizi di igiene e prevenzione epidemiologica, medici di base, medicina del lavoro, servizi di continuità assistenziale) mettendoli in grado di isolare e contenere tempestivamente catene di contagio e focolai (approccio **test, trace & treat**).
2. **Strumenti data-driven per il monitoraggio**: potenziare il sistema di sorveglianza epidemiologica con la capacità di integrare **molteplici sorgenti di dati** in **dashboard analitiche** e modelli di previsione per il **monitoraggio** di indicatori dell’epidemia nelle fasi successive.
3. **Open data sanitari e Intelligenza Artificiale**: rendere disponibili per l’analisi **i dati clinici e radiologici** così da ingaggiare centri di ricerca su progetti di **big data analytics e intelligenza artificiale** e avanzare la conoscenza sulla malattia mediante modelli predittivi e esplicativi del decorso clinico dei pazienti Covid-19.
4. **Sistemi informativi sanitari centrati sul paziente**: garantire sistemi informativi che connettano i servizi della medicina sul territorio con l’assistenza ospedaliera, gestendo tutto il **ciclo clinico ed extra-clinico del paziente Covid** (diagnostica, assistenza a domicilio, quarantena, assistenza ospedaliera)

Il report del 14/04/2020, esteso con **integrazioni** sviluppate nelle settimane successive, è stato discusso il 29/04/2020 con il Ministro dell’Innovazione Paola Pisano dalla quale sono arrivate indicazioni per investigare ulteriormente la fattibilità dei punti 2 e 3.

In questo periodo le comunità scientifiche dell’epidemiologia mondiale hanno ampiamente discusso quali segnali monitorare per affinare modelli predittivi dell’evoluzione dell’epidemia. In Italia, i dati epidemiologici forniti giornalmente dalla protezione civile, per quanto affetti da errori e limitati ad una risoluzione regionale o provinciale, sono stati utilizzati per raffinare in itinere e

validare le tante variazioni di modelli epidemici proposti. I progressivi interventi di lockdown hanno anche confermato lo stretto legame tra mobilità e diffusione del Covid-19[1][2][3]. Alcune analisi preliminari hanno utilizzato dei proxy dei comportamenti di mobilità reale, come ad esempio le stime macroscopiche di mobilità fornite da Google[4]. E' invece possibile ed auspicabile utilizzare proxy più accurate e a maggiore risoluzione, quali quelle fornite dai dati della telefonia mobile[5], dove la stima della mobilità può essere inferita dalle tracce (call data record CDR e XDR) lasciate dagli utenti, sulla base delle quali gli operatori della telefonia mobile possono rilasciare giornalmente i flussi aggregati origine-destinazione, sia inter- che intra-comunali per l'intero paese, in modalità completamente anonima. E' importante sottolineare che, tra le numerose variabili utili ad affinare modelli predittivi, oltre ai dati epidemiologici provenienti dal sistema sanitario, i dati di mobilità da telefonia sono gli unici che possono essere resi disponibili a cadenza giornaliera. Mentre i dati degli operatori telefonici catturano in unico segnale gli spostamenti delle persone con qualsiasi mezzo di trasporto, è ovviamente auspicabile anche la disponibilità di dati veicolari e di uso dei mezzi pubblici di trasporto.

I modelli predittivi sono stati uno strumento fondamentale per valutare diverse ipotesi per la i riapertura. Il documento presentato dal Comitato Tecnico Scientifico[6] il 26 Aprile mostra la valutazione dei rischi di diffusione epidemica del Covid-19 in diversi scenari di rilascio del lockdown. Gli scenari, anche sulla base di dati Istat ed INAIL, esplorano le dimensioni che impattano la diffusione del virus in una società stratificata per età come la nostra al momento della ripresa di attività economiche e sociali, variando la percentuale di individui che ritornano attivi nei vari settori e tenendo conto dell'impatto, specifico per ogni settore, che questo ha sui contatti nei luoghi di lavoro e nelle comunità. Per ogni scenario, i modelli hanno anche consentito di stimare il numero di riproduzione effettivo Rt e la prevalenza di casi severi al picco dell'epidemia.

Questo approccio aiuta a definire una strategia di riapertura graduale affiancata da un sistema di sorveglianza basato su una ampia serie di indicatori che catturino l'evoluzione dell'epidemia, ma anche la resilienza dei sistemi sanitari ad eventuali picchi, e la loro capacità di tracciare l'epidemia stessa (in particolare di effettuare tamponi). La strategia definita dal Ministero della Salute nel decreto del 30/04/2020[7] relativo agli strumenti di monitoraggio per la fase 2 descrive tre famiglie di indicatori e le relative sorgenti informative:

- indicatori di processo sulla capacità di monitoraggio;
- indicatori di processo sull'accertamento diagnostico, l'indagine e la gestione dei contatti;
- indicatori della trasmissione del virus e della tenuta dei servizi sanitari.

[1] Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures ProfileMarino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. PNAS first published April 23, 2020 https://doi.org/10.1073/pnas.2004978117

[2] The effect of human mobility and control measures on the COVID-19 epidemic in China MUG Kraemer, CH Yang, B Gutierrez, CH Wu, B Klein, DM Pigott, Science 368 (6490), 493-497

[3] The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak M Chinazzi, JT Davis, M Ajelli, C Gioannini, M Litvinova, S Merler, Science 368 (6489), 395-400

[4] Google: Covid-19 Community Mobility Report https://www.gstatic.com/covid19/mobility/2020-04-26_IT_Mobility_Report_en.pdf (26 Aprile 2020).

[5] CNR-UNIPI-WINDTRE (P. Bonato, P. Cintia, F. Fabbri, D. Fadda, F. Giannotti, P. Lopalco, S. Mazzilli, M. Nanni, L. Pappalardo, D. Pedreschi, F. Penone, S. Rinzivillo, G. Rossetti, M. Savarese, L. Tavoschi): Mobile phone data analytics against the COVID-19 epidemics in Italy flow diversity and local job markets during the national lockdown. https://arxiv.org/abs/2004.11278

[6] Valutazione di politiche di riapertura utilizzando contatti sociali e rischio di esposizione professionale. Stefano Merler, Daniele Checchi, conferenza stampa CTS, 26 Aprile, 2020

[7] Emergenza COVID-19: attività di monitoraggio del rischio sanitario connesse al passaggio dalla fase 1 alla fase 2. A di cui all'allegato 10 del DPCM 26/4/2020. Decreto Ministero della Salute 30.04.2020

Anche i documenti elaborati della Associazione degli Epidemiologi[8] convergono sull'insieme di indicatori da monitorare per valutare il rischio epidemico delle regioni, presentano una accurata analisi delle sorgenti informative disponibili, incluso il sistema di sorveglianza epidemica dell'Istituto Superiore di Sanità[9] e dettagliano la tipologia di dati che le regioni devono produrre. Su **provenienza, qualità e disponibilità dei dati della sorveglianza epidemica è importante uno sforzo di chiarezza e trasparenza**, volto a superare le criticità emerse nella fase 1 e a rendere accessibili le informazioni, in formati adeguati e con tempestività, anche ai ricercatori e agli innovatori per sviluppare conoscenze e soluzioni sulle sfide poste dall'epidemia. Le migliori esperienze internazionali[10] mostrano chiaramente la validità e la fattibilità di **approcci moderni di "open science" attraverso infrastrutture dati che offrano modalità controllate e sicure per abilitare ricerche scientifiche e sperimentazione di soluzioni innovative**, coniugando la salvaguardia della privacy e delle finalità etiche con l'accesso sia a dati aggregati che a micro-dati anonimizzati sui percorsi di malattia dei pazienti Covid-19.

1. Ecosistema integrato "data-driven" per il monitoraggio del rischio sanitario

Il decreto del Ministero della Salute del 30/04/2020 definisce le attività di monitoraggio del rischio sanitario connesso al passaggio alla fase 2 della gestione dell'epidemia. Sulla base di indicatori e criteri, il decreto disegna a tutti gli effetti un (**eco**)**sistema integrato "data-driven" di supporto alle decisioni**, la cui affidabilità e tempestività è fondamentale per il rilascio progressivo delle restrizioni della fase 1 in sicurezza, evitando la risorgenza del contagio. Un tale sistema **non è solo composto di dati e di software** per calcolare indicatori, produrre visualizzazioni, realizzare analisi statistiche e stimare modelli predittivi, ovvero di una dashboard (cruscotto) di analisi dati per il monitoraggio degli indicatori. Esso è anche (se non soprattutto) composto di **soggetti portatori di interesse** sia nella **produzione dei dati** che alimentano il sistema, sia **nell'uso di informazioni ed evidenze** da esso generate, in base alle quali effettuare **valutazioni di rischio e prendere decisioni di rilascio o inasprimento delle restrizioni sociali** (quando, dove, come). L'efficacia del sistema integrato è fortemente legata all'identificazione e al coinvolgimento dei soggetti citati, sia a livello locale (regionale, comunale) che centrale (Ministero della Salute e Comitato Tecnico Scientifico, Istituto Superiore di Sanità, Conferenza delle Regioni, altri Ministeri in grado di contribuire dati e strumenti d'analisi, ISTAT, ...). Questo indirizzo è apertamente mirato a rispondere all'emergenza informativa su due piani collegati:

- iniziando una **profonda innovazione organizzativa** centrata soprattutto sulle persone, ovvero sullo sviluppo di **team di "data scientist"** che, sia a livello centrale che locale, siano **portatori della cultura del dato** e facilitino le diverse fasi della **filiera di trasformazione dei dati in conoscenza utile alla evidence-based policy;**
- iniziando ad **"uscire i dati per la ricerca scientifica"**, ovvero interfacciando il sistema integrato dei dati con la comunità scientifica in modo tempestivo e sicuro, in un chiaro contesto di **governance dei dati e della progettualità "open science"**, per massimizzare la capacità di risposta alle tante domande aperte sulla malattia Covid-19, sia a livello epidemiologico che clinico (anche in relazione all'iniziativa sui dati clinici suggerita nella sez. 2 di questo documento).

[8] Il contributo dell'epidemiologia per orientare le attività di sanità pubblica ed assistenziali durante la fase 2 della epidemia Covid-19 in Italia. Analisi e proposte della Associazione Italiana di Epidemiologia https://www.epiprev.it/documento-aie-22-aprile-2020

[9] Allegato n. 03_DOC AIE 22 aprile 2020, I dati e gli indicatori della sorveglianza regionale e dell'Istituto Superiore di Sanità. A cura di Paola Angelini e Serena Broccoli

[10] Ad esempio l'iniziativa OpenSafely in UK, "*a new secure analytics platform for electronic health records in the NHS, created to deliver urgent results during the global COVID-19 emergency. It is now successfully delivering analyses across more than 24 million patients' full pseudonymised primary care NHS records, with more to follow shortly. All our analytic software is open for security review, scientific review, and re-use.*" https://opensafely.org/

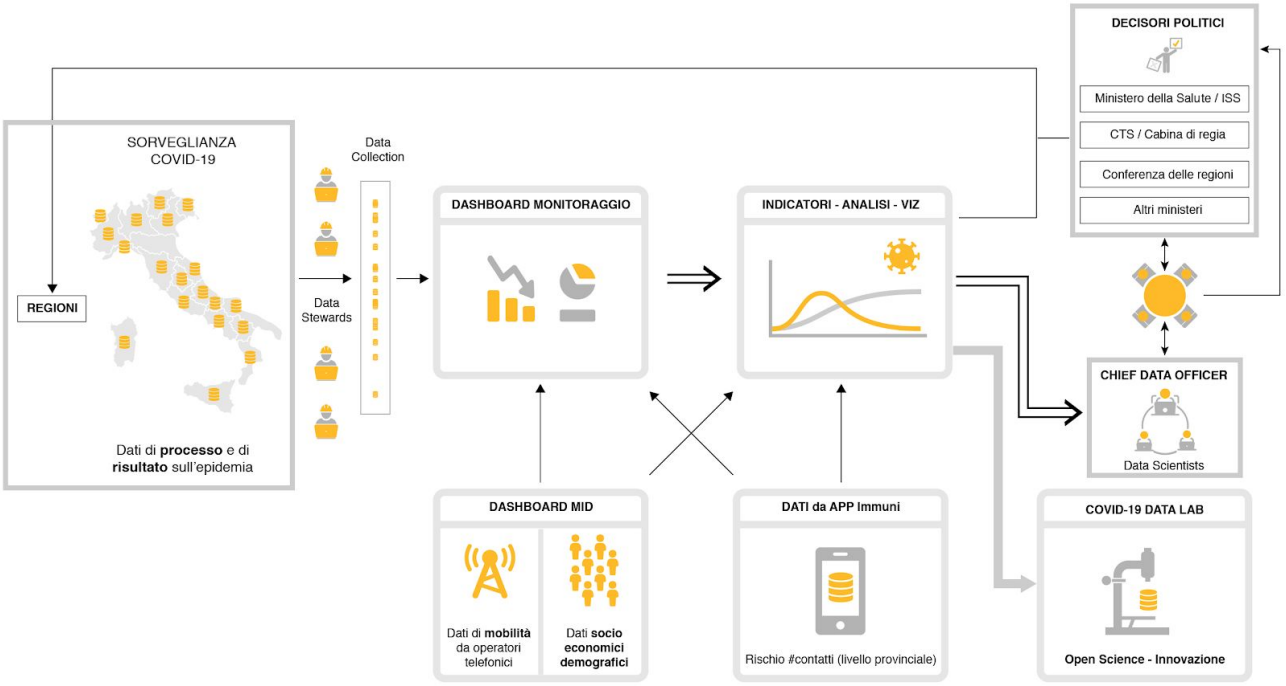


Figura: Architettura dell'ecosistema integrato "data-driven" per il monitoraggio

Aspetti tecnici e organizzativi chiave:

- Flussi di raccolta dei dati epidemiologici (sistemi informativi regionali, dipartimenti di prevenzione, sorveglianza integrata Covid di ISS, Ministero della Salute, Protezione Civile)
- Team distribuito di "data stewards" di raccordo con i sistemi regionali per la raccolta dati
- Flussi di raccolta dati sull'impatto sul sistema sanitario
- Flussi di raccolta dei dati di mobilità (operatori telefonici, altre sorgenti)
- Popolamento di dati demografici, socio-economici e ambientali (Ministeri, ISTAT, INAIL, ...)
- Controllo di qualità ed integrazione dati
- Costruzione degli indicatori di monitoraggio
- Cruscotto di navigazione e visualizzazione degli indicatori
- Strumenti di analisi statistica e modelli predittivi
- Sistemi di alert in base alle soglie di rischio sanitario
- Interfacce rivolte ai diversi stakeholder centrali e locali
- Team di "data scientist" coordinati da un CDO per il raccordo con i decisori
- Interfacce "Data Lab" per Open science & innovation

In riferimento anche al lavoro di disegno di una infrastruttura di dati svolto dal sottogruppo 2 (Infrastrutture e data collection), il sottogruppo 7 (Big Data & AI for policy) raccomanda che il Ministero della Salute, il Ministero per l'Innovazione Tecnologica e la Digitalizzazione e la Conferenza delle Regioni si impegnino in un progetto congiunto rivolto a realizzare tempestivamente il sistema integrato "data-driven" dispiegando le risorse umane e materiali necessarie allo scopo - in particolare, le figure di "**data steward**" sul territorio per facilitare il raccordo con i sistemi regionali per la raccolta dati, e le figure di "**data scientist**", coordinate da un Chief Data Officer, per realizzare le analisi dati a supporto della policy e facilitare il raccordo con gli stakeholder e i decisori a livello centrale e locale. L'urgenza e la necessità di un intervento di questo tipo giustifica una partnership fra i due ministeri e le regioni che sappia valorizzare le competenze diffuse nel sistema sanitario in un contesto di innovazione digitale, che aiuti a superare la frammentazione organizzativa, e che favorisca l'integrazione dei dati e lo sviluppo di

modelli di interpretazione del fenomeno epidemico. Il sottogruppo 7 ritiene che questo approccio sistemico sia l'unico in grado di garantire efficacia nel governo di fenomeni complessi come la pandemia e supportare il coordinamento degli interventi, specialmente non farmaceutici, posti in essere per il contenimento, e rimane a disposizione per affiancare l'indirizzo e l'accompagnamento del progetto delineato.

2. Datathon e infrastruttura dati per dati clinici dei pazienti Covid-19

La malattia Covid-19 è ancora non del tutto compresa anche nella sua evoluzione clinica. Sono state riscontrate presentazioni cliniche variabili e presumibilmente influenzate da multipli fattori, paziente-dipendenti (età, comorbidità, genotipo) e ambiente-dipendenti (area geografica) che necessitano una maggiore comprensione. Per questo motivo la ricerca scientifica deve esplorare la multifattorialità della malattia ed identificare i modelli di paziente che siano utili allo sviluppo di terapie appropriate. In questo contesto la diagnostica per immagini ha un ruolo importante, trovandosi a valutare sul campo, nei percorsi COVID, migliaia di pazienti con impegno polmonare, ma anche multiorgano (cardiovascolare, neurologico, gastrointestinale).

Le immagini radiologiche rappresentano **una espressione fenotipica della malattia**, un bagaglio enorme di informazioni digitali che possono essere utilizzate anche attraverso l'ausilio di nuove soluzioni di **Intelligenza Artificiale,** quali ad esempio il **Deep Learning.** I dati visuali radiologici arricchiti con le informazioni cliniche e ancor di più con i profili genomici (genotipi) dei pazienti stessi, offrono l'opportunità per una più approfondita costruzione di conoscenza della malattia. Una stretta interazione tra l'esperienza medica e la capacità di apprendimento e di modellizzazione dell' intelligenza artificiale moderna, rese possibili dalla potenza computazionale attualmente disponibile, aprono la strada ad una **intelligenza collaborativa esperto-macchina**, non solo ai fini diagnostici ma anche ai fini prognostici e a supporto delle scelte terapiche adeguate (si veda in proposito la posizione della Società Italiana di Radiologia Medica SIRM[11]).

In questo contesto, il gruppo di lavoro, interagendo con il Laboratorio CINI "Artificial Intelligence and Intelligent Systems", che aggrega la comunità scientifica AI Italiana, e la Società Italiana per la Radiologia Medica ha elaborato una proposta progettuale[12] di carattere interministeriale (Ministero dell'Innovazione e la Trasformazione Digitale, Ministero della Salute, Ministero dell'Università e della Ricerca, Istituto Superiore di Sanità) che vada nella direzione di creare una infrastruttura di dati clinici dei pazienti Covid-19 nello spirito di Open Science, ovvero creando le condizioni per un accesso sicuro e responsabile ai dati, nel rispetto della normativa GDPR, che fornisca un terreno fertile per l'avanzamento della conoscenza fondato su criteri di fondo:

- L'uso dei dati, compresa la condivisione, l'aggregazione e l'analisi dei dati, per la risposta al Covid19 deve riferirsi a chiare esigenze espresse dalle autorità sanitarie pubbliche, e non per altri scopi.
- L'uso dei dati deve essere conforme alle leggi esistenti e rispettare i principi delle migliori pratiche di governance dei dati.

[11] Neri, E., Miele, V., Coppola, F. et al. Use of CT and artificial intelligence in suspected or COVID-19 positive patients: statement of the Italian Society of Medical and Interventional Radiology. Radiol med (2020). https://doi.org/10.1007/s11547-020-01197-9

[12] Medicina ed Intelligenza Artificiale Italiana: Una Alleanza per il per il COVID. Fosca Giannotti, Dino Pedreschi, Mauro Grigioni, Task Force MdI) Rita Cucchiara, CINI Lab Artificial Intelligence and intelligent Systems Roberto Grassi, Emanuele Neri Società Italiana per la Radiologia Medica

- I dati devono essere aggregati alla più bassa risoluzione possibile, mantenendo l'utilità desiderata.
- L'uso dei dati deve essere trasparente, inclusivo e salvaguardato da conseguenze indesiderate.

A questo scopo si possono investigare svariati paradigmi:

- piattaforme analitiche che permettano un approccio "*move the algorithm to the data*" che, invece di spingere i dati dalle aziende regionali verso un server centrale, punta ad eseguire l'algoritmo nei "repository regionali" per poi condividere i risultati ottenuti localmente (tabelle di sintesi, distribuzioni, modelli) combinandoli con i risultati da altri repository. Esempi in questa direzione sono l'esperienza inglese https://opensafely.org/ realizzata in pochissimo tempo proprio per rispondere a Covid-19, o anche l'approccio OPAL (OPen ALgorithms, https://www.media.mit.edu/projects/opal-health/overview) di MIT[13] e la più recente Mobility Data Network https://www.covid19mobility.org/
- tecniche di anonimizzazione dei dati che mappano le features dei singoli casi su valori a maggiore sicurezza e interpretabilità, mediante l'utilizzo di ontologie di dominio che possono facilitare l'accesso ai dati e la loro condivisione
- tecniche di condivisione centralizzata mediante protocolli di rete di *onion routing* (si veda il progetto TOR https://www.torproject.org) che permettono a sorgenti diverse di inviare dati ad un server centrale senza che questo possa in alcun modo risalire alla sorgente mittente. Tale approccio potrebbe essere utile nel caso le risorse computazionali necessarie per esecuzioni deep learning non fossero disponibili nei repository locali, e purché i dati da condividere siano anonimi.

Obiettivo ultimo è la creazione di una infrastruttura dati distribuita sul territorio nazionale che integri immagini e dati clinici di pazienti COVID-19 positivi (fenotipo), e di dati genomici dei pazienti sottoposti al test RT-PCR (genotipo).

La proposta prevede:

A. La raccolta immediata e controllata di dati medici fenotipici e genotipici riguardo il COVID attraverso una chiamata diretta e volontaria da parte dell'Istituto Superiore di Sanità di strutture ospedaliere capaci di fornire dati ed annotazioni.
B. La sperimentazione di tecniche di Intelligenza artificiale che garantiscano una efficacia nella diagnosi, nella prognosi, nella analisi quantitativa dei dati, uniti ad un grado di spiegabilità e di affidabilità compatibile con le direttive Italiane ed europee ed italiane
C. La creazione di una infrastruttura federata coordinata dall'Istituto Superiore di Sanità che, mediante i paradigmi più recenti di accesso ai big data in modo sicuro, e compatibilmente con le direttive sulla privacy e gli aspetti legali, possa mantenere il grado attuale di controllo, di proprietà ed indipendenza delle strutture sanitarie italiane e permettere nel contempo una elaborazione efficace del contenuto, attraverso strumenti digitali
D. La adozione di un modello di raccolta e di gestione di dati che segua le normative già definire da SIRM dal Febbraio 2020 [6] per predisporre un referto strutturato specifico per il COVID che seguendo norme standard DICOM possano diventare uno standard per la refertazione e poi per il supporto alla analisi digitale e agli strumenti di apprendimento automatico e di AI.

Siamo consapevoli che tale proposta ha un carattere strutturale, naturalmente ancorato al Ministero Salute, ma riteniamo che i punti A e B potrebbero essere realizzati nella forma di un

[13] Hardjono, T., Kim, A., & Pentland, A. (2020). 7. Health IT: Algorithms, Privacy, and Data. In Building the New Economy. Retrieved from https://wip.mitpress.mit.edu/pub/a56wpt24

datathon da realizzarsi entro l'estate, mirato anche alla promozione ed esemplificazione pratica di un'importante innovazione tecnologica e culturale di condivisione tempestiva e sicura del dato a fini scientifici. Troppo spesso si interpreta "Open Science" come Open Data e quindi ci si barrica dietro ai vincoli etico-legali, non considerando invece varietà di accesso ai dati FAIR (Findable, Accessible, Interoperable and Responsible) accessi on-site (gli esperimenti si fanno dove sono i dati e si prendono in carica le responsabilità); e tecnologie di Machine Learning che non prevedono lo spostamento dei dati (Federated Learning).

Proposta datathon

Il datathon è un passo preliminare che può aiutare a capire meglio la potenzialità, l'utilità e la fattibilità dell'infrastruttura. A questo fine il disegno del datathon dovrà seguire le seguenti linee.

1. **CHALLENGE**:
    a. **una domanda clinica/medica**: estrarre misure verificabili sui dati sia con metodi di tipo supervisionato che semi-supervisionato (dalla segmentazione, alla correlazione, al clustering per similarità, alla classificazione) che leghino evidenze fenotipiche (su immagini e su dati clinici) alla gravità del paziente e al decorso nel tempo della malattia, a fini prognostici.
    b. **una domanda tecnologica**: è possibile costruire tali metodi senza spostare i dati? oppure quali tecniche di anonimizzazione usare?

2. **DATI**: immagini e dati clinici di pazienti COVID-19 positivi (fenotipo), e di dati genomici dei pazienti sottoposti al test RT-PCR (genotipo) **anonimizzati** utilizzando l'ontologia medica definita da SIRM. Per questo si pensa alla possibilità di correlare dati multi-omici di pazienti costruendo "Digital patient twins". I dati per la sperimentazione verticale saranno necessariamente non completi (alcuni Digital twin avranno solo dati immagini o clinici in quanto le mappe genomiche sono complesse da ottenere rispetto ad altri referti), parzialmente annotati (si spera annotati con il referto strutturato, ma che in alcuni casi potrà essere ricostruito) e in altri non validati; i dataset, per poter misurare specificità e sensibilità ed altre misure comprensibili all'esperto, dovranno contenere anche dati non COVID sia di altre malattie respiratorie sia di pazienti sani.

3. **FORMATO**: si propone di usare il format utilizzato con successo dalla Infrastruttura di ricerca Europea SoBigData[14] che ha realizzato datathon in vari ambiti in molti paesi Europei: calcio, fake-news, apps, wellbeing. Si veda ad esempio: https://www.sobigdata-soccerchallenge.it/trento/ Il formato è in due tempi, il primo serve a selezionare i gruppi (qualificazione) che competono e nel secondo si da l'accesso ai dati ai gruppi selezionati che hanno un tempo designato per produrre il risultato.
    a. QUALIFICAZIONE: I team sottopongono una proposta e la valutazione da parte di un comitato scientifico adotterà appropriati criteri quali composizione ed esperienza multi-disciplinare del team, qualità della proposta di ricerca.
    b. FASE FINALE: I team selezionati potranno accedere ai dati di un ospedale che, con l'approvazione del proprio comitato etico, ha accettato di partecipare. I componenti dei team, dietro assunzione di responsabilità, potranno accedere ai dati anonimizzati di "immagini e dati clinici di pazienti COVID-19 positivi (fenotipo), e di

[14] www.SoBigData.eu

dati genomici dei pazienti sottoposti al test RT-PCR (genotipo)", ed avranno un tempo limitato, ed esempio una settimana, per presentare i risultati.

c. SELEZIONE DEI VINCITORI: in base a criteri quali robustezza e accuratezza valutata su test set di altri ospedali, comprensibilità e fiducia dei modelli proposti da parte dei medici.

4. **RISULTATI ATTESI:** non sono solo di tipo scientifico, ma anche individuazione delle problematiche di standardizzazione degli strumenti, delle linee guida per la trasparenza e l'architettura della infrastruttura.
5. **PIANO DI AZIONE SUGGERITO**: istituire un comitato organizzatore (CO) inter-istituzionale (MID, Min Salute, IIS, SIRM, AIIS-CINI) con il compito di individuare gli ospedali disponibili, individuare i dataset utilizzabili, il formato dei dati, le annotazioni, le anonimizzazioni, le modalità di processing etc., definire il quadro etico-legale e costruire il bando. Il CO dovrà nominare un Comitato scientifico che avrà il compito di effettuare la selezione dei partecipanti. Il CO potrà avvalersi della infrastruttura europea di ricerca SoBigData.eu in collaborazione con AIIS-CINI per assicurare le appropriate risorse computazionali e di accesso ai dati e la gestione dei partecipanti.

3. Indicazioni tecniche per l'implementazione del flusso descritto nel decreto del Ministero della Salute del 30/4/2020[15]

Nel seguito sono riportate alcune considerazioni e indicazioni per l'implementazione del sistema software di integrazione ed analisi dati in riferimento al decreto del Min Salute. Il decreto definisce 21 indicatori divisi in tre gruppi

1. indicatori della capacità di monitoraggio;
2. indicatori sulla capacità di accertamento diagnostico, indagine e gestione dei contatti;
3. indicatori sulla stabilità di trasmissione e tenuta dei servizi sanitari

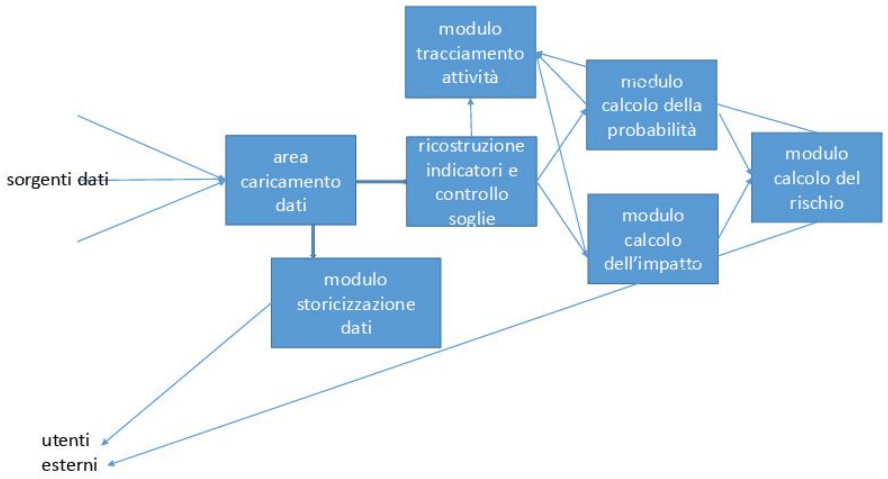
Si tratta di indicatori sintetici (21 numeri, alcuni assoluti, altri percentuali) per la cui collezione ed utilizzo si ritiene fondamentale un sistema caratterizzato da

- semplicità
- tracciabilità delle operazioni
- elevata modularità.

I dati dovrebbero essere inviati esclusivamente in forma elettronica con un formato testuale predefinito (*e.g.,* file csv). Dal decreto si può assumere che i fornitori dei dati siano in numero limitato (soprattutto l'Istituto Superiore di Sanità, Protezione Civile, Regioni, lo stesso Ministero della Salute). Per ognuno di questi soggetti dovrebbe essere definita un'identità digitale ad hoc con vincolo di firma (digitale) di tutti i contenuti inviati.

Data la mole ridotta dei dati, si può ipotizzare un'organizzazione snella del sistema schematizzata nel seguente modo

[15] a cura di Massimo Bernaschi, IAC-CNR



Il flusso dei dati tra i moduli è basato sulla definizione di un formato testuale organizzato per coppie chiave-valore che permette il trasferimento sia in memoria sia utilizzando file di appoggio, ispezionabili a fini di controllo e per un eventuale intervento diretto.
Le sorgenti dati, all'atto del caricamento (effettuato previa autenticazione), ricevono una conferma con *digest* del file inviato[16].
Il cuore del sistema è costituito dal modulo di *ricostruzione indicatori e controllo delle soglie.* Questo modulo assembla, ove necessario, i singoli dati (ad esempio per quelli forniti dalle Regioni) richiesti per costruire l'indicatore complessivo ed effettua il controllo di qualità registrando l'eventuale mancanza di dati, la non conformità al formato stabilito, *etc.* Si ritiene utile dotare il modulo di un sistema di *alert* che avvisi automaticamente sia le sorgenti dati sia i responsabili del sistema integrato delle situazioni che determinano una riduzione del livello di qualità dei dati e quindi dell'intero servizio.
Utenti esterni possono accedere tutti i dati caricati (direttamente dal modulo di storicizzazione[17]) ed il risultato finale, ovvero il rischio calcolato sulla base dei dati e delle procedure definite nei moduli di calcolo della probabilità e dell'impatto.
Mentre i moduli per la ricostruzione degli indicatori, per il calcolo della probabilità, dell'impatto e del rischio non sono direttamente accessibili dall'esterno, la loro descrizione ed il software con cui sono realizzati sono resi pubblici, per permettere a chiunque di ricostruire i risultati ottenuti. Per lo scopo della riproducibilità, sarebbe opportuno che fosse reso disponibile oltre all'indicatore 3.2 (Rt) anche tutti i dati utilizzati per la sua determinazione (database ISS elaborato dalla Fondazione Bruno Kessler). Il completamento di questo sistema dovrebbe avvenire prima che i dati inizino ad essere raccolti e quindi gestiti in modalità dettate dall'emergenza.

[16] Si potrebbe pensare di caricare il *digest* su una *blockchain* ma probabilmente è eccessivo.
[17] Eventualmente previa registrazione

4. Gruppo di lavoro

La stesura di questo rapporto è stata curata da:

- Dino **Pedreschi** Univ. Pisa, Direttore Phd Data Science, Direttore KDD Lab, coordinatore gruppo 7 "Big Data & AI for policy" task force MID (Computer Science, Big data, AI)
- Fosca **Giannotti** ISTI-CNR, Pisa, coordinatrice scientifica Research Infrastructure SoBigData.eu (Computer Science, Data Science, Big data, AI)
- Francesca **Chiaromonte** Scuola Superiore S.Anna, Pisa, coordinatrice scientifica Dipartimento di Eccellenza EMbeDS, Economics and Management in the era of Data Science (Statistica)
- Paolo **Vineis** Imperial College, London, Environmental Epidemiology (Epidemiologia)
- Massimo **Bernaschi** IAC-CNR, Roma, (Mathematics, Computer Science)
- Mauro **Grigioni**, Istituto Superiore di Sanità, Direttore Centro Nazionale Tecnologie Innovative in Sanità Pubblica (Bioingegneria)
- Serafino **Sorrenti** responsabile Agenda Digitale Regione Sicilia, Monitoraggio Migrazione per Ministero Salute (Economia, Innovazione digitale)
- Luca **Ferretti**, Univ. Oxford, Physics, Statistical Genetics and Pathogen Dynamics Epidemiology (Epidemiologia computazionale, Fisica)
- Stefano **Merler**, FBK Trento, Direttore Unità di Ricerca Dynamical Processes in Complex Societies (Epidemiologia computazionale)
- Elio **Mungo**, Poleecy (Startup, Imprenditoria)
- Paolo **De Rosa** Dipartimento Trasformazione Digitale, Responsabile Task force data-driven anti Covid-19, Min. Innovazione (Computer Science, Innovazione digitale)

5. Riferimenti

1. "Cosa fare per interrompere la catena dei contagi" a cura dell'Associazione Italiana di Epidemiologia (AIE), Scienza in rete, Aprile 2020 https://www.scienzainrete.it/articolo/cosa-fare-interrompere-catena-dei-contagi/associazione-italiana-di-epidemiologia-aie/2020
2. Il contributo dell'epidemiologia per orientare le attività di sanità pubblica ed assistenziali durante la fase 2 della epidemia Covid-19 in Italia. Analisi e proposte della Associazione Italiana di Epidemiologia https://www.epiprev.it/documento-aie-22-aprile-2020
3. CNR-UNIPI-WINDTRE (P. Bonato, P. Cintia, F. Fabbri, D. Fadda, F. Giannotti, P. Lopalco , S. Mazzilli , M. Nanni, L. Pappalardo, D. Pedreschi, F. Penone, S. Rinzivillo, G. Rossetti, M. Savarese, L. Tavoschi): Mobile phone data analytics against the COVID-19 epidemics in Italy flow diversity and local job markets during the national lockdown. ArXiv https://arxiv.org/abs/2004.11278
4. Google: Covid-19 Community Mobility Report https://www.gstatic.com/covid19/mobility/2020-04-26_IT_Mobility_Report_en.pdf (26 Aprile 2020).
5. Suppression of COVID-19 outbreak in the municipality of Vo', E Lavezzo, E Franchin, C Ciavarella et al. Italy. medRxiv, 18-04-2020. https://doi.org/10.1101/2020.04.17.20053157
6. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, Christophe Fraser https://science.sciencemag.org/content/early/2020/04/09/science.abb6936
7. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures ProfileMarino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. PNAS first published April 23, 2020 https://doi.org/10.1073/pnas.2004978117
8. The effect of human mobility and control measures on the COVID-19 epidemic in China MUG Kraemer, CH Yang, B Gutierrez, CH Wu, B Klein, DM Pigott, Science 368 (6490), 493-497

9. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. M Chinazzi, JT Davis, M Ajelli, C Gioannini, M Litvinova, S Merler, Science 368 (6489), 395-400
10. Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak. Brennan Klein, Timothy LaRock, Stefan McCabe, Leo Torres, Filippo Privitera, Brennan Lake, Moritz U. G. Kraemer, John S. Brownstein, David Lazer, Tina Eliassi-Rad, Samuel V. Scarpino, Matteo Chinazzi, and Alessandro Vespignani. Report: https://uploads-ssl.webflow.com/5c9104426f6f88ac129ef3d2/5e8374ee75221201609ab586_Assessing_mobility_changes_in_the_United_States_during_the_COVID_19_outbreak.pdf
11. Valutazione di politiche di riapertura utilizzando contatti sociali e rischio di esposizione professionale. Stefano Merler, Daniele Checchi, conferenza stampa CTS, 26 Aprile, 2020
12. Emergenza COVID-19: attività di monitoraggio del rischio sanitario connesse al passaggio dalla fase 1 alla fase 2. A di cui all'allegato 10 del DPCM 26/4/2020. Decreto Ministero della Salute 30.04.2020
13. Hardjono, T., Kim, A., & Pentland, A. (2020). 7. Health IT: Algorithms, Privacy, and Data. In Building the New Economy. Retrieved from https://wip.mitpress.mit.edu/pub/a56wpt24
14. Neri, E., Miele, V., Coppola, F. et al. Use of CT and artificial intelligence in suspected or COVID-19 positive patients: statement of the Italian Society of Medical and Interventional Radiology. Radiol med (2020). https://doi.org/10.1007/s11547-020-01197-9
15. T. Hardjono and A. Pentland, "MIT Open Algorithms," in Trusted Data - A New Framework for Identity and Data Sharing, T. Hardjono, A. Pentland, and D. Shrier, Eds. MIT Press, 2019, pp. 83–107.

6. Allegati

1. Primo Report SGdL 7 "Big Data & AI" del 14.04.2020, disponibile online: https://github.com/taskforce-covid-19/documents/blob/master/sgdl_7_Big_Data_AI_Policies/sgdl7_analisi_stakeholder_e_raccomandazioni.pdf

2. Proposta Medicina & AI (Datathon e Infrastruttura Dati Covid) a cura di SGdL7, SIRME (Soc. Ital. Radiologia Medica), ISS, Lab AIIS (Artificial Intelligence & Intelligent Systems) del CINI.

Medicina ed Intelligenza Artificiale Italiana: Un'alleanza per il COVID

"Explainable AI for COVID data"

Rita Cucchiara[1], Fosca Giannotti[1], Roberto Grassi[2], Mauro Grigioni[3], Emanuele Neri[2] , Dino Pedreschi[1]

(1) CINI Lab Artificial Intelligence and intelligent Systems, (2) SIRM - Società Italiana di Radiologia Medica e Interventistica (3) ISS - Istituto Superiore di Sanità

L'emergenza COVID-19 sta stimolando la comunità scientifica internazionale a proporre sperimentazioni finalizzate a migliorare la diagnosi, l'inquadramento clinico-diagnostico e lo sviluppo di terapie mirate nei pazienti COVID-19 positivi.
Sono state riscontrate presentazioni cliniche variabili e presumibilmente influenzate da multipli fattori, paziente-dipendenti (età, comorbidità, genotipo) e ambiente-dipendenti (area geografica) che necessitano una maggiore comprensione. Per questo motivo la ricerca scientifica deve esplorare la multifattorialità della malattia ed identificare i modelli di paziente che siano utili allo sviluppo di terapie appropriate.
In questo contesto l**a diagnostica per immagini** ha un ruolo importante, trovandosi a valutare sul campo, nei percorsi COVID, migliaia di pazienti con impegno polmonare, ma anche multiorgano (cardiovascolare, neurologico, gastrointestinale).

Le immagini radiologiche vengono quindi a rappresentare **una espressione fenotipica** della malattia, un bagaglio enorme di informazioni digitali (da accedere secondo i paradigmi definiti in ambito big data) che possono essere utilizzate anche attraverso l'ausilio di nuove soluzioni di **Intelligenza Artificiale.**
I dati visuali radiologici non sono sufficienti: questi dati devono però essere aggregati con **le informazioni cliniche e ancor di più con i profili genomici (genotipi)** dei pazienti stessi, per una più completa costruzione di conoscenza con stretta interazione tra l'esperienza medica, la potenza computazionale dei calcolatori attuali e la capacità di apprendimento e di modellizzazione della intelligenza artificiale moderna.
Gli strumenti di Intelligenza Artificiale (AI), frutto della ricerca informatica dell'ultimo decennio, che associano al *medical imaging* tradizionale la potenza dell'apprendimento automatico e delle reti neurali profonde (*Deep Learning*) possono costituire un valido ausilio per una **intelligenza collaborativa esperto-macchina**, non solo ai fini diagnostici ma anche ai fini prognostici e a supporto delle scelte terapiche adeguate.
I nuovi sistemi di Intelligenza Artificiale hanno già' trovato un validissimo ruolo in molti contesti medici (per un survey sugli ultimi risultati nel mondo tratti dalle più' importanti convegni sull'AI e sulla medicina si veda [1]); recentemente sono stati dichiarati utili anche nell'analisi del COVID-19 (si vedano le dichiarazioni in termini di sensitività e specificità di AI Alibaba Cloud, testato – in base ai report- in Cina[2]).

Malgrado le recenti dichiarazioni e gli evidenti successi delle ricerca e dei prodotti di AI in tutti i campi del sapere, compresi quelli legati alla salute, **esistono ancora problemi aperti su come e se i tool esistenti possano essere adottati come strumenti affidabili nel panorama europeo ed italiano.** Non e' detto che dati utilizzati per l'apprendimento siano consistenti con i dati che derivano dai macchinari italiani, che la annotazione segua le direttive in uso nel nostro paese, né che i risultati in termini diagnostici siano direttamente replicabili.
Nel contempo molte soluzioni di ricerca che provengono da tutto il mondo e anche dai laboratori italiani propongono **soluzioni open source** di grande interesse a supporto del mondo medico, sia per la pre-elaborazione dei dati multispettrali 2D e 3D, per il riconoscimento di zone di interesse, per la diagnosi, il confronto quantitativo, nonché' di tecniche di estrazione di conoscenza multimodale, impegnando dati visuali, dati testuali (informazioni di referti in keywords o linguaggio naturale) e dati clinici, sensoriali e contestuali differenti.
I nuovi sistemi di AI offrono quindi potenzialmente molte soluzioni da esplorare, che attualmente però non possono essere neppure verificate, o applicate in Italia senza un valido coordinamento nazionale, un contatto stretto con i sistemi sanitari regionali, e senza una sinergia organizzativa tra le competenze mediche ed informatiche del nostro paese.

E' estremamente strategico che in questo momento di emergenza, e in relazione alle prossime fasi di supporto continuo all'assistenza e alla terapia, si possa sperimentare l'utilizzo delle tecnologie esistenti e i nuovi risultati della ricerca AI spesso prodotti direttamente dalla ricerca italiana: sono strumenti che già si stanno testando ovunque nel mondo, ed in modo sporadico e non coordinato anche in Italia; e' necessario un progetto unico, coordinato a livello del sistema Paese.

Anche seguendo le direttive suggerite dalla commissione Europea (si veda il recente White paper AI [3]), n**asce la necessità di testare, validare, adattare e spesso riprogettare un software** che abbia le caratteristiche di apprendere dai dati e produrre conoscenza in collaborazione con l'essere umano e che possa essere dotato del grado **di affidabilità, spiegabilità e riproducibilità** che ci si attende da strumenti di uso in ambito critico come la salute.
Servono strumenti di apprendimento automatico, di elaborazioni di dati visuali e multimodali ma anche strumenti di spiegabilità (explainable AI) e di validazione applicati sui dati COVID italiani.

E' necessario quindi ed in tempi rapidissimi, strutturare un progetto di ricerca, **ma direttamente applicabile** che porti a contatto al comunità di esperti medici e la comunità di esperti di intelligenza Artificiale in Italia per lavorare su software prodotto in Italia e validato da laboratori italiani, con dati che derivano dalla esperienza del nostro paese e certificati dagli ospedali già' fortemente coinvolti nella emergenza.
Un primo passo sarebbe già la preparazione di un Datathon che produca a brevissimo termine una conoscenza di cosa può essere realizzato in Italia.
Si propone quindi un progetto **che abbia risultati a breve termine e che abbia l'ambizione di rappresentare, iniziativa unica in Italia, una infrastruttura a lungo termine** seguendo i dettami

che provengono dagli esperti, e nel caso specifico dalla **Società Italiana di Radiologia Medica ed Interventistica (SIRM)** [7,8,9,10,11] e coadiuvati dagli esperti informatici che, in più di 50 laboratori, sono rappresentati dal **Lab di Intelligenza Artificiale e sistemi Intelligenti del consorzio Interuniversitario Nazionale di Informatica, (CINI Lab AIIS) [4],** collegati in modo diretto con le comunità Europee ed internazionali. Tra queste si citano tra le altre le iniziative in essere a livello europeo quali i progetti *Human-AI,* ed il recente *HumanAI-Net* e *SoBigData*, anche per il supporto alle infrastrutture per l'accesso a dati in modo distribuito e *DeepHealth* per la creazione di librerie AI europee per la comprensione di immagini e dati 3D sulla salute [6]. Ad essi si affiancano i diversi progetti Europei a cui partecipano i medici ed i radiologi Italiani, in collaborazione con le strutture ospedaliere del nostro paese.

La proposta prevede

a) L**a raccolta immediata e controllata di dati medici fenotipici e genotipici** riguardo il COVID attraverso una chiamata diretta e volontaria da parte dell'Istituto Superiore di Sanità' di strutture ospedaliere capaci di fornire dati ed annotazioni.

b) **La sperimentazione di tecniche di Intelligenza artificiale che garantiscano una efficacia nella diagnosi, nella prognosi, nella analisi quantitativa dei dati, uniti ad un gradi di spiegabilità e di affidabilità** compatibile con le direttive Italiane ed europee ed italiane (Questo e' stato dichiarato prioritario anche dai documenti italiani sulla strategia per l'Intelligenza artificiale, recentemente licenziati dal Governo Italiano [4] e le proposte che derivano dal CINI lab AIIS [5])

c) La **creazione di una infrastruttura federata** coordinata dall'Istituto Superiore di Sanità che possa mediante i paradigmi più' recenti di accesso ai big data in modo sicuro, e compatibile con le direttive sulla privacy e gli aspetti legali, mantenere il grado attuale di controllo e di proprietà ed indipendenza delle strutture sanitarie italiane e permettere nel contempo una elaborazione efficace del contenuto, attraverso strumenti digitali

d) **La adozione di un modello di raccolta e di gestione di dati che segua le normative già' definire da SIRM** dal Febbraio 2020 [7] per predisporre un referto strutturato specifico per il COVID che seguendo norme standard DICOM possano diventare uno standard per la refertazione e poi per il supporto alla analisi digitale e agli strumenti di apprendimento automatico e di AI.

Obiettivo ultimo **e' la creazione di una rete federata di biobanche integrate, immagini e dati clinici di pazienti COVID-19 positivi (fenotipo), e di dati genomici dei pazienti sottoposti al test RT-PCR, (genotipo) distribuita sul territorio nazionale**:

Esse dovranno avere requisiti e caratteristiche simili, criteri uniformi di raccolta e classificazione della banca dati, indice di casi raccolti quotidianamente aggiornati, accessibili a ricercatori italiani. Una simile biobanca integrata costituirà una risorsa unica ad oggi nel panorama scientifico internazionale e possibile a realizzarsi solo in presenza di una sistema sanitario universale, gratuito e dotato di infrastrutture di alto livello, quale quello italiano.

Tale risorsa potrà favorire lo sviluppo di algoritmi di intelligenza artificiale del nostro paese, promuovere anche l'economia italiana nel mondo IT e soprattutto ottenere risultati mirati alla migliore comprensione delle caratteristiche multifattoriali della malattia e determinare quanto una malattia sia condizionata da fattori genetici, ambientali o di stili di vita.

Riferimenti bibliogarfici

[1]Z. Ryan Shi, C.Wang, f.Fang Artificial Intelligence for Social Good: A Survey Carnegie mellon University Jan 2020 https://arxiv.org/abs/2001.01818

[2] https://www.alibabacloud.com/solutions/ct-image-analytics

[3] "White Paper of Artificial Intelligence A European approach to excellence and trush" ttps://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[4] "Proposte per la Startegia Italiana in Intellgienza Artificiale " Gruppo di Lavoro MISE https://www.mise.gov.it/index.php/it/strategia-intelligenza-artificiale/contesto

[5] - "AI for Future Italy" - Lab CINI for Artificial Intelligence and Intelligent Systems , Febbraio 2020

[6] E Tartaglione, CA Barbano, C Berzovini, M Calandri, M Grangetto Unveiling COVID-19 from Chest X-ray with deep learning: a hurdles race with small data arXiv preprint arXiv:2004.05405 2020

[7] F.Coppola, L. Faggioni, D. Regge, A.Giovagnoni, R. Golferi, C Bibbolino · V. Miele, E. Ner, R. Grassi Artifcial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey La Radiologia Medica , Feb. 2020

[8] E.Neri, V. Miele, F. Coppola, R. Grassi "Use of CT and artificial intelligence in suspected or COVID‐19 positive patients: statement of the Italian Society of Medical and Interventional Radiology, La Radiologia Medica April 2020

[9] Neri E, Coppola F, Miele V, Bibbolino C, Grassi R. Artificial intelligence: Who is responsible for the diagnosis? [published online ahead of print, 2020 Jan 31]. *Radiol Med*. 2020;10.1007/s11547-020-01135-9. doi:10.1007/s11547-020-01135-9

[10] Neri, E., de Souza, N., Brady, A. *et al.* What the radiologist should know about artificial intelligence – an ESR white paper. *Insights Imaging* 10, 44 (2019). https://doi.org/10.1186/s13244-019-0738-2

[11] Neri E, Regge D. Imaging biobanks in oncology: European perspective. *Future Oncol*. 2017;13(5):433‐441. doi:10.2217/fon-2016-0239