PROPOSTA DI UNA INDAGINE A CAMPIONE PER UNA STIMA AFFIDABILE DEI PARAMETRI FONDAMENTALI DELLA EPIDEMIA DA SARS-CoV-2

1. PREMESSA E MOTIVAZIONE

Anche alle soglie della prevedibile conclusione della fase di emergenza, per il controllo della diffusione del SARS-CoV-2 e la verifica del livello raggiunto del tasso di immunizzazione si impone un'analisi accurata delle fonti dei dati. È importante, infatti, disporre di basi informative pienamente affidabili sulle quali fondare la stima dei principali parametri epidemiologici e la previsione dell'evoluzione per poter suggerire interventi di politica sanitaria e di riavvio delle attività economiche e sociali.

Nella situazione di grande incertezza sulla dinamica di un fenomeno nuovo e per molti aspetti ignoto risulta cruciale la sua misura quanto più possibile accurata. Ciò è rilevante tanto nella fase ascendente dei contagi, quanto in quella discendente. L'incertezza rimarrà, ed è ineludibile, ma il livello di imprecisione nel monitoraggio dell'evoluzione dell'epidemia può e deve essere ridotto e mantenuto sotto controllo.

I dati sui tamponi finora raccolti sono basati su un campionamento 'di convenienza' che ha privilegiato l'esame dei casi che manifestavano sintomi. Come è noto, un campione di convenienza non consente di produrre una stima probabilistica non distorta e con un prefissato livello di accuratezza. È stato messo in evidenza da diversi studi come, sulla base dei dati raccolti finora, risulti fortemente sottostimato il numero di contagiati essendo quasi completamente assenti quelli asintomatici, con una conseguente sovrastima del tasso di letalità. Anche la somministrazione massiva di tamponi o di test sierologici in alcune regioni non rappresenta un campione probabilistico.

La conoscenza chiara delle dinamiche in atto ci sembra ineludibile per le misure da assumere da parte dei responsabili politici e sanitari oltre che per la consapevolezza e i comportamenti da adottare da parte della popolazione. E' necessario un impianto di rilevazione che consenta di produrre stime non distorte e di fare confronti statisticamente significativi nel tempo e tra diverse aree geografiche, tenendo conto dei differenti contesti economici, demografici, sociali, ambientali e culturali.

Inoltre, non ci si può limitare a misurare in modo accurato il numero dei contagi, dei ricoveri e dei loro esiti, i risultati dei tamponi e degli esami sierologici, ma occorre fare luce sulle caratteristiche individuali, famigliari e di contesto che possano favorire o ostacolare l'infezione e valutare correttamente l'effetto degli interventi adottati per modificare l'evoluzione del fenomeno. Un lavoro da fare insieme, epidemiologi, virologi, responsabili di strutture sanitarie, esperti di progettazione di indagini e di modelli matematici e statistici ed esperti di valutazione di politiche pubbliche.

Per ricavare stime affidabili e quindi realmente utili riteniamo sia indispensabile, compatibilmente con l'impegno di forze attualmente dispiegato per il controllo

dell'emergenza, progettare e realizzare un protocollo di osservazione a campione riferito all'intera popolazione italiana, seguendo un disegno statisticamente rigoroso. Sul campione deve essere effettuato il tampone e/o l'esame sierologico, a seconda dello specifico obiettivo del monitoraggio.
Questo consentirebbe di valorizzare l'informazione preziosa che viene raccolta da parte delle strutture sanitarie secondo le indicazioni delle istituzioni preposte, integrandola all'interno di un quadro complessivo che contenga la stima accurata delle variabili d'interesse in domini temporali e spaziali individuati opportunamente.

L'uso integrato di diverse fonti, fra le quali - in primo luogo - quelle amministrative nel corso dell'accoglienza e della cura delle persone che si rivolgono al sistema sanitario, quelle statistiche progettate per misurare con accuratezza la diffusione del contagio e/o dell'immunizzazione, quelle fornite da nuove fonti idonee a tracciare gli spostamenti delle persone e i loro contatti - rappresenta una necessità per poter fornire dati di qualità, non distorti e non fuorvianti, sui quali fondare decisioni importanti per la salute pubblica. È una sfida metodologica, tecnologica e organizzativa alla quale la comunità degli statistici, e certamente degli statistici ufficiali, può dare un contributo utile.
L'indagine campionaria proposta rappresenta, dunque, un tassello della costruzione di un sistema informativo integrato che consenta di rispondere a una pluralità di obiettivi:

a) monitorare l'evoluzione del fenomeno e stimarne le modalità di diffusione;
b) stimare correttamente il numero effettivo dei contagiati e della popolazione immune, i tassi di asintomaticità, guarigione e letalità;
c) valutare l'effetto dell'introduzione o rimozione di misure restrittive;
d) orientarne i tempi nelle diverse fasi dell'evoluzione dell'epidemia;
e) prepararsi ad affrontare in modo informato le prossime prevedibili epidemie/pandemie.

L'uscita dall'epidemia in corso avrà una durata relativamente estesa e le regole per la ripresa progressiva delle attività economiche e sociali, in piena sicurezza, di porzioni di territori, tipologie di nuclei famigliari e di imprese impongono una conoscenza il più possibile nitida dei rischi relativi; soprattutto per acquisire la necessaria fiducia e condivisione da parte dei cittadini e degli operatori economici riguardo a misure che saranno prevedibilmente di carattere selettivo.

Fiducia e condivisione devono potersi fondare su una base informativa solida.

* * * * *

Si propone un protocollo di osservazione che consenta di fornire stime statisticamente significative della dimensione delle diverse componenti della popolazione, definite in relazione all'epidemia da SARS-CoV-2. Uno strumento dinamico di monitoraggio progettato in modo da poter essere opportunamente calibrato sia nella fase di crescita dei contagi, fornendo la loro stima secondo diverse categorie di gravità, sia nella fase di decrescita, con stime estese ai parametri della progressiva immunizzazione della popolazione. Stime tutte con affidabilità definita a livello temporale e territoriale.

Con riferimento alla popolazione infetta, a partire da quella per la quale sia stato accertato l'avvenuto contagio (conclamati), l'obiettivo è quello di stimare la popolazione contagiata ma non ancora diagnostica.

Invece, per la stima nel corso del tempo delle dimensioni della popolazione secondo la relazione passata o presente con il virus, è necessario riferirsi a un campione rappresentativo dell'intera popolazione.
Ai fini della procedura proposta vengono individuati, in ciascun momento temporale, due sottoinsiemi di individui che denominiamo: *Gruppo d'interesse A* e *Gruppo d'interesse B*:
Gruppo d'interesse A. È il sottoinsieme che comprende le persone il cui stato di infezione sia conclamato (che possono essere ricoverati o in quarantena coatta) e tutte le persone che hanno avuto contatti con loro, risalenti fino a 14 giorni prima; si tratta, quindi, dell'insieme di persone prevedibilmente contagiate e non soltanto di quelle per le quali il contagio sia stato già accertato.
Gruppo d'interesse B. È il sottoinsieme complementare di tutte le persone non entrate in contatto con quelle del *Gruppo A.* Esso comprende le persone sane, le persone guarite, le persone immuni, le persone asintomatiche, e inoltre le persone in fase di incubazione per le quali i sintomi si manifesteranno successivamente, nell' arco di massimo 14 giorni.
Le stime relative ai due sottoinsiemi possono essere ottenute sulla base di un'osservazione continua nel tempo e seguendo due metodologie distinte, entrambe basate sul campionamento *indiretto*, lo stesso che si usa per la stima di popolazioni *rare* o *elusive*.
La proposta è stata formalizzata da un punto di vista metodologico e il testo in lingua inglese è pubblicato alla pagina web http://arxiv.org/abs/2004.0606. È importante sottolineare che le stime statistiche degli aggregati d'interesse sono effettuate con il concorso di ambedue i campioni ed è stato dimostrato analiticamente che esse godono di proprietà importanti: la non distorsione e l'efficienza.
La proposta è stata valutata anche in termini di fattibilità organizzativa, stimandone i costi e i tempi necessari per la realizzazione. Si sottolinea che solamente una piena sponsorizzazione dell'iniziativa da parte delle autorità sanitarie può consentirne l'impianto. Ad esempio, nel caso di monitoraggio a livello regionale, la titolarità e responsabilità dell'indagine dovrebbe essere attribuita alla Regione, considerata la sua competenza in materia di provvedimenti per la salute pubblica.
Si riporta qui di seguito una descrizione delle fasi nelle quali si articolano i due campioni attraverso i quali si propone di impiantare il sistema di sorveglianza dell'epidemia.

2. CAMPIONAMENTO PER IL GRUPPO A

Lo schema di campionamento è finalizzato alla stima della popolazione infetta senza sintomi, o in fase di incubazione, ed è continuo nel tempo. Una volta definito il periodo di sorveglianza, lo schema si articola nelle seguenti fasi:

A1. *selezione per ogni unità di tempo (giorno, settimana, altro) di un campione di conclamati (persone risultate infette)*. In una fase di forte attenuazione dei contagi si può considerare il totale dei nuovi infetti. La selezione avviene tra i ricoverati presso strutture ospedaliere o in quarantena presso la propria abitazione. Da un punto di vista organizzativo ci si può avvalere dell'elenco dei contagiati che si sono riferiti a presidi sanitari in una finestra temporale fissata opportunamente; oppure, in mancanza dell'elenco, si può utilizzare un campione di strutture ospedaliere (unità di primo stadio) e quindi arrivare alla selezione dei contagiati (unità di secondo stadio).

A2. *ricostruzione di tutti i contatti, risalenti fino a 14 giorni prima, delle persone selezionate al punto A1*. La ricostruzione dei contatti può essere effettuata ripercorrendo le principali reti sociali, secondo la usuale distinzione in reti primarie (parenti, amici, vicini di casa etc.), reti secondarie (colleghi di lavoro, commessi di negozi alimentari, farmacisti etc.), reti formali (medici, forze dell'ordine, operatori sanitari etc.), reti informali (comunità, gruppi di aggregazione etc.). Per semplificare questa fase, si potrebbero utilizzare apposite *app* disponibili su *smartphone* per il tracciamento delle persone;

A3. *somministrazione dei tamponi e/o dei test sierologici*, a seconda delle finalità che ci si prefigge, a un campione di persone che hanno avuto contatti con quello selezionato al punto A1.

Il campionamento nelle fasi A1 e A2 dovrebbe essere realizzato tenendo conto delle dimensioni spaziale e temporale.

- ✓ *Campionamento spaziale*. L'uso di tecniche di campionamento spaziale permette di avere un campione *panel* ottimale di strutture sanitarie, selezionate in base alla loro dimensione e altre caratteristiche; ottimale in termini sia di copertura territoriale sia logistici;
- ✓ *Campionamento temporale.* Nelle strutture sanitarie selezionate attraverso il campionamento spaziale, l'individuazione del campione di persone conclamate delle quali ricostruire i contatti può essere realizzata per intervalli di tempo definiti di accesso alla struttura (ricovero, altro). Ad esempio, potrebbe essere selezionata una persona tra i conclamati che in un intervallo di tempo prefissato hanno avuto accesso alla struttura sanitaria, ripetendo la selezione per ogni intervallo temporale.

Come già sottolineato, la dimensione del campione di conclamati e dell'unità di tempo di riferimento per la loro selezione è definita in funzione del livello di contagio. Ai fini della stima statistica del numero di persone contagiate in un dato dominio *territoriale* (territorio nazionale/specifica area geografica come, ad esempio, una regione) potrebbe essere sufficiente coinvolgere nell'indagine per circa 1.000 persone tra i contatti delle persone conclamate sulle quali effettuare i tamponi (ed eventualmente il test sierologico). Questa dimensione campionaria, assicurerebbe una stima affidabile con un errore relativo di campionamento intorno al 5% qualora la proporzione dei contagiati

tra i contatti della persone infette (conclamati) fosse intorno al 25%. L'errore relativo diminuirebbe aumentando la dimensione campionaria.
Ipotizzando la ricostruzione di circa 25 contatti per ogni conclamato, le persone alle quali somministrare i 1.000 tamponi potrebbero essere individuate:

- ✓ selezionando circa 200 persone conclamate (di cui si ricostruirebbero 5.000=200x25 contatti);
- ✓ campionando una quota di circa il 20% dei contatti di cui sopra.

Ai fini sanitari, per la prevenzione del contagio, sarebbe opportuno seguire tutti i contagiati. Tuttavia, da un punto di vista statistico, per ottenere stime di elevata qualità sul numero di persone contagiate nello specifico dominio di studio pianificato questo non è necessario.

3. CAMPIONAMENTO PER IL GRUPPO B

Per la stima del numero di contagiati al di fuori dei contatti delle persone conclamate deve essere selezionato un campione *panel* di individui da seguire in modo continuo nel tempo. Le fasi operative sono le seguenti:

B1. il panel viene sottoposto al tampone o altri test regolarmente (ad esempio una volta a settimana o bimensilmente);
B2. se una persona del panel risulta positiva sono ricostruiti tutti i suoi contatti nei precedenti 14 giorni;
B3. Il tampone e gli altri esami vengono somministrati anche a un campione di questi contatti.

Il numero di persone da coinvolgere nel campione *panel* potrebbe essere di circa 1.000 (per circa 1.200 tamponi/esami) per un dato dominio territoriale, garantendo una stima con un errore relativo di campionamento inferiore al 10% qualora la proporzione dei contagiati per il target fosse intorno al 10%. L'errore relativo diminuirebbe aumentando la dimensione campionaria.

Il *panel* potrebbe essere formato con una metodologia in due fasi, in modo da selezionare le persone mediante un *pre-screening* che può fare riferimento a quattro categorie di soggetti.
Un primo gruppo composto da persone ad alto rischio (ad esempio, con età superiore a 50 o 60 anni) diviso in due sottogruppi: persone con attività intensa e con contatti ripetuti nel corso della giornata e persone con vita più ritirata e pochi contatti interpersonali.
Un secondo gruppo composto da persone a basso rischio, distinto nei medesimi due sottogruppi.

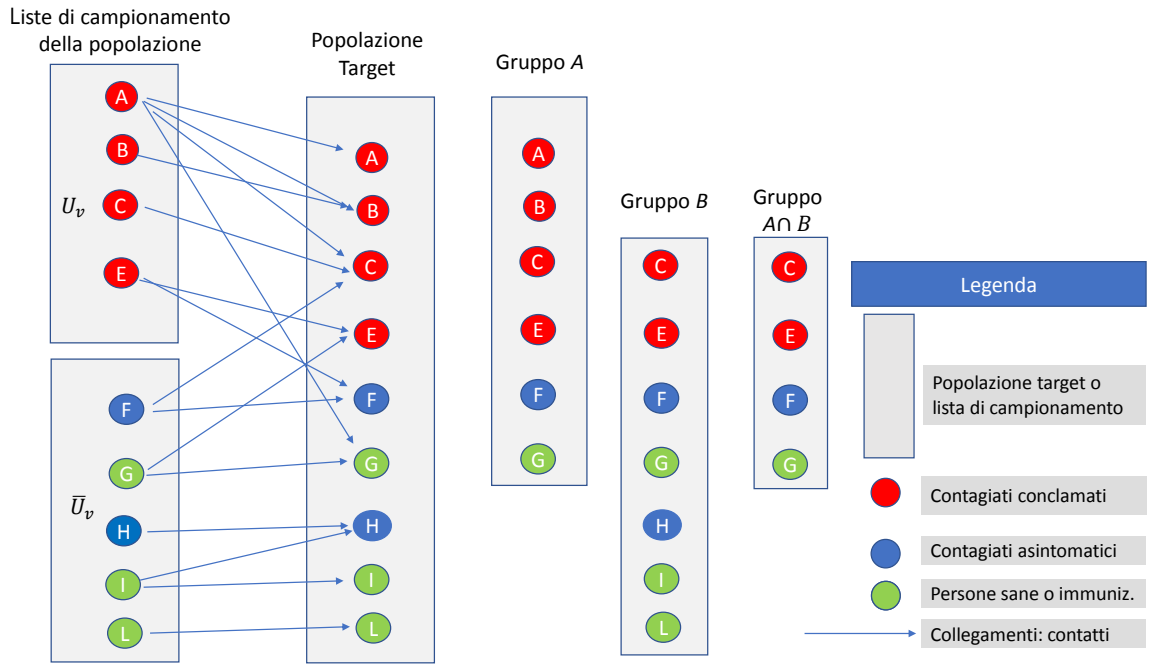
Le stime statistiche degli aggregati d'interesse sono effettuate con il concorso di ambedue i campioni (la stima delle intersezioni tra i due gruppi A e B può essere ottenuta sovrapponendo i campioni). È stato dimostrato analiticamente che esse godono di due importanti proprietà: la non distorsione e l'efficienza.

In conclusione, il protocollo di osservazione che si propone è in grado di fornire stime statisticamente affidabili su variabili significative dell'epidemia, uno strumento dinamico di monitoraggio, con la possibilità di accentuare l'osservazione sul primo campione nel caso si avessero segnali del riacutizzarsi del contagio o dell'arrivo di una seconda ondata e sul secondo campione, il *panel*, per valutare, ad esempio attraverso esami sierologici, gli andamenti nel corso della fase 2.

(aggiornato al 26 aprile 2020)

Giorgio Alleva, ordinario di statistica, Sapienza Università di Roma, ex presidente dell'Istat.
Giuseppe Arbia, ordinario di statistica economica, Università Cattolica di Roma, esperto di analisi spaziale dei dati.
Piero Demetrio Falorsi, ex direttore della direzione metodologica dell'Istat, esperto di disegni di campionamento.
Guido Pellegrini, ordinario di statistica economica, Sapienza Università di Roma, presidente della Commissione di garanzia dell'informazione statistica, esperto di valutazione di politiche pubbliche.
Alberto Zuliani, professore emerito di statistica, già presidente dell'Istat.

Campioni e popolazioni d'interesse nel disegno di campionamento proposto



Legenda. *Popolazione target*: insieme delle persone infette conclamate, infette asintomatiche, sane (comprensive delle immunizzate). *Liste di campionamento*: U_v= conclamati selezionati per la ricostruzione dei contatti; $\bar{U}_v$= campione panel. *Gruppo A*: conclamati selezionati e campione di loro contatti. *Gruppo B*: panel e campione di contatti di quanti del panel risultati infetti.