Frauds in scientific research and how to possibly overcome them

Erik Boetto, Davide Golinelli, Gherardo Carullo, Maria Pia Fantini

ABSTRACT
Frauds and misconduct have been common in the history of science. Recent events connected to the COVID-19 pandemic have highlighted how the risks and consequences of this are no longer acceptable. Two papers, addressing the treatment of COVID-19, have been published in two of the most prestigious medical journals; the authors declared to have analysed electronic health records from a private corporation, which apparently collected data of tens of thousands of patients, coming from hundreds of hospitals. Both papers have been retracted a few weeks later. When such events happen, the confidence of the population in scientific research is likely to be weakened. This paper highlights how the current system endangers the reliability of scientific research, and the very foundations of the trust system on which modern healthcare is based. Having shed light on the dangers of a system without appropriate monitoring, the proposed analysis suggests to strengthen the existing journal policies and improve the research process using new technologies supporting control activities by public authorities. Among these solutions, we mention the promising aspects of the blockchain technology which seems a promising solution to avoid the repetition of the mistakes linked to the recent and past history of research.

INTRODUCTION
Frauds and misconducts have been common in the history of science. Recent events connected to the COVID-19 pandemic have highlighted how the risks and consequences of this are no longer acceptable, especially during a global health emergency. The time has come to review the entire editorial policy system, especially the process of evidence creation.

The objective of this paper is to highlight how the current system endangers the reliability of scientific research and the very foundations of the trust system on which modern healthcare is based. Having shed light on the dangers of a system without appropriate monitoring, both for the scientific community and society as a whole, we propose to strengthen the existing journal policies and improve the research process using new technologies supporting control activities by public authorities. Among these solutions, we mention the promising aspects of the blockchain technology which seems a promising solution to avoid the repetition of the mistakes linked to the recent and past history of research.

AN EXCESSIVE AMOUNT OF SCIENTIFIC LITERATURE MAY LOWER THE OVERALL QUALITY OF RESEARCH
The number of papers published on COVID-19 has reached tens of thousands, and is still growing. Scientists, policymakers and physicians are facing increasing difficulty in finding articles relevant to their activities, while also evaluating the quality of the scientific evidence provided.

COVID-19 pandemic is giving us a glimpse of how scientific research might evolve in the future: researchers want to share their work in a quick, open-access fashion, in order to receive immediate feedback and recognition by their peers. Preprint servers allow them to do so, as shown by their increased popularity: almost 4000 COVID-19 papers have been submitted to medRxiv alone as of June 2020.1

While this ‘torrent of preprints’ surely contributed to raise the level of the COVID-19 ‘sea of literature’, the major flow is represented by published articles in refereed journals.2 The peer review process should designate the researches deserving of publication; however, scientific journals must ensure that all the potential discoveries on COVID-19 are given a chance to be divulged in a timely fashion. In order to cope with the overwhelming amount of submissions, many journals had to decrease the time required to evaluate and publish a manuscript. A recent study—ironically, a preprint—shows that 14 medical journals have decreased by almost 50% the time required on average to publish a COVID-19 paper by reducing the number of days required for the peer review.3

These issues raised concerns about the quality of COVID-19 publications, the spread of misinformation and lack of scientific integrity, which are the foundations of the so-called ‘infodemic’.

A BRIEF INTRODUCTION TO RECENT EVENTS: THE SURGISPHERE SCANDAL
Recent events show us that even those journals that have always been considered among the best in their sector are not immune to COVID-19 infodemic.

Two papers have been recently published in two of the most prestigious medical journals only to be retracted a few weeks later. The first one—published on The New England Journal of Medicine (NEJM) on 1 May 2020—addressed the lack of harmful effects of ACE inhibitors and angiotensin receptor blockers in patients with COVID-19;4 the other—published on Lancet on 22 May 2020—focused on the potential risks of using hydroxychloroquine (HCQ) or chloroquine as COVID-19 treatment.5

These two papers share three authors (Professor Mandeep R Mehra, Dr Sepan S Desai and Dr Amit N Patel)4 5 and also the source of the data analysed: Surgisphere, a data analytics company founded by Dr Desai in 2007. Both studies were based on Surgisphere electronic health records (EHR) data, which
collected data of tens of thousands of patients, coming from hundreds of hospitals from up to six continents.

Such statements raised scepticism among the readers of said journals: how was it possible for a small private company to obtain access to such an amount of international data? This, along with concerns on the methods and results of said papers, prompted NEJM to issue an Expression of Concern (EOC) on 3 June, asking the authors to provide evidence that the data are reliable; Lancet also issued a similar EOC. Two days later, both journals retracted the respective publications, since ‘the authors were not granted access to the raw data’ and ‘Surgisphere would not transfer the full dataset […] for analysis’, corroborating the suspicions of data fraud and manipulation.45

In the research addressing the use of HCQ or chloroquine in patients with COVID-19, the authors stated that this treatment ‘was associated with decreased in-hospital survival and an increased frequency of ventricular arrhythmias when used for treatment of COVID-19’.5 A randomised UK trial, Randomised Evaluation of COVID-19 Therapy, reported a similar lack of benefit from HCQ treatment, but found no evidence of toxicity to the heart.6 The heart toxicity reported by Mehra et al contributed to a global halt of the ongoing research on HCQ as a treatment for COVID-19, prompting even the WHO itself to take a ‘temporary pause’ of the HCQ trial, which was later resumed after the retraction.7

A third paper was coauthored by Mehra, Desai and Patel, once again backed by Surgisphere data. Submitted on 6 April to a preprint server, the paper—recently removed by the authors and no longer available—reported a reduced mortality in patients with COVID-19 treated with ivermectin, an antiparasitic drug. Carlos J Chaccour, from Barcelona Institute for Global Health,8 says the preprint has influenced the healthcare of some Latin American countries: on 2 May a white paper encouraged the inclusion of ivermectin in Peru national guidelines for COVID-19.9 The example of Peru was followed by Bolivia and Paraguay, where an unsustainable demand of ivermectin brought to the use of the veterinary formulations of the drug, sold both regularly and through the black market.8

Since this research had been divulged only as a preprint, no journal had to issue an EOC followed by an official retraction. The legacy of this fraudulent research is still alive and will continue to cause harm through false sense of security, possible side effects and shortage of the drug for its appropriate applications.

‘The damage to public health continues, fuelled by unbalanced media reporting and an ineffective response from government, researchers, journals, and the medical profession[…]’. This statement, fitting to the ivermectin/COVID-19 situation, was issued in 2011 by the British Medical Journal, quickly after former doctor Andrew J Wakefield incrimination.10

On 28 February 1998, Wakefield published a study on the Lancet reporting a link between measles, mumps and rubella vaccination and a syndrome of autism and bowel disease in 12 children.11

In the years that followed, many epidemiological studies found no evidence of this link;10 however, the study was retracted only 12 years later, in 2010, when the claims that children were consecutively referred and that investigations were approved by the local ethics committee had been proven to be false.10

Between 1998 and 2010, Wakefield’s unethical misconduct caused a long-lasting damage, through diminished vaccination rates and general distrust of the healthcare authorities, sentiments that are still lingering today, even during the pandemic.12

WEAKNESS IN THE DATA PROCESS

The Surgisphere scandal has brought to light the criticalities in the world of scientific publishing, but what happened is not an anomaly. To understand how this kind of misconduct has been possible, we must consider the processing of scientific data (eg, clinical data), from when it is created to when it ends up in the guidelines for clinical practice. While this process is country by country dependent, there are some common points.13

The data are generated by one individual (eg, a patient) in a healthcare setting (figure 1). An ‘intermediary’ (eg, a researcher) usually collects, sorts it and analyses it. Then, the results of this analysis must be shared and reported through a divulger (ie, scientific journal), in order to be used by the rest of the scientific community and beyond. Finally, the public or private authorities (WHO, Ministry of Health, other hospitals) receive the information and put into practice the new evidence that returns as a benefit to the end user (again the patient).

This should be a virtuous circle (figure 1). However, what has been brutally highlighted by the recent events is that there is a number (unknown) and a type (unknown) of intermediaries who claim to collect and analyse huge volumes of digitised real-world data. What determines an increased risk of process failure is not knowing who these subjects are, their interests and the way in which they collect, store, analyse and share data. As previously outlined, data collected during a research phase can be altered ex post for many reasons.14 In particular, any subjects involved in the research can be led to review some of the data already collected, in order to confirm the expected results. At best, such data manipulation can determine the grave consequences described before. In worst-case scenarios, the falsification of data leads to the commercialisation of potentially harmful products, with serious consequences for patients’ health, related ethical and legal problems and also economic and social issues.

PUBLIC (DIS)TRUST IN THE DATA PROCESS: UNDERMINING TRUST IN HEALTHCARE AND SCIENTIFIC RESEARCH

In modern societies trust of the general public in healthcare and scientific research is based, inter alia, on the credibility of the data process in biomedical scientific research described above. Such a process is usually quite difficult for the non-technical population to understand. The general population is thus normally informed by scientific news sources and the mass media, which have the important task of translating the scientific results achieved by researchers into a language intelligible to everyone.

The results of the above described process therefore gain credibility in the eyes of the population both because of the credibility of the researchers, and in light of the actual outcomes, including patient satisfaction.15

However, if the results actually produced are not in line with expectations, or worse, turn out to be based on fraudulent data, the confidence of the population in the aforementioned process is likely to be weakened. Where this malfunction is occasional or in any case sporadic, the responsibility for errors might only be attributed to individual persons. Conversely, where a certain dysfunction is repeated, and occurs independently of the subjects involved, the distrust might shift towards the process itself. This second hypothesis is particularly dangerous because in this case about the effects of data fraud and manipulation.
Current controversy

the solution to the problem perceived by the population is no longer the simple replacement of the subjects in charge. In such cases, the effectiveness of the process as a whole might be questioned.

This is what could happen in relation to the phenomenon of falsification, or in any case manipulation, of health data. As explained above, numerous cases have already occurred in which it was ascertained that the data on which certain scientific evidence was based were in whole or in part false, or in any case were not entirely accurate. Therefore, the repetition of these events can have the consequence of undermining the credibility of the process as a whole. The population, considered as a whole, does not have the knowledge and the means to tell apart reliable and unreliable scientific news. As a consequence, repeated instances in which it has been demonstrated the unreliability of well-established sources have the effect of leaving the general public with the doubt of what information can actually be trusted.16 The more this happens, the more grows the distrust of the population towards scientific news sources and the mass media and, consequently, towards healthcare and scientific research processes.

To prevent this loss of confidence, we must evaluate how this dysfunction can be resolved. In order to avoid a complete reform of modern scientific processes, we can rather (1) strengthen the existing journal policies, ensure they are really adhered (see the Strengthening existing journal policies section), and (2) support editorial activities with new tools (see the Blockchain technology to support the data process in scientific research section), to make up for their current shortcomings.

STRENGTHENING EXISTING JOURNAL POLICIES

The most immediate causes of research misconduct potentially relate to journal policies regarding the deposition of data and substandard reviews. Many journals should ensure the reviews are of a high standard, particularly during a devastating pandemic. However, this does not always happen.

In the retracted *Lancet* paper5 the authors state that they used ‘de-identified data obtained by automated data extraction from inpatient and outpatient EHRs, supply chain databases, and financial records’. Similarly, the authors implicated in the *NEJM* article4 stated that they were unable to validate the primary data sources because they ‘were not granted access to the raw data and the raw data could not be made available to a third-party auditor’.

This should have raised some queries and the nature of the said data repositories should have been investigated. Almost all journals require investigators to submit a data sharing statement and register a data sharing plan when registering a trial. However, many of them have not yet formulated a policy on which types of data sharing plans are acceptable. For example, the *Lancet*’s author guidelines on this matter state that ‘for reports of research other than clinical trials, data sharing statements are encouraged but not required’. This lax approach has probably contributed to the problem. Hence, it seems that a strengthening of existing editorial tools and their constant and uniform adoption could have helped mitigate the risk in most cases of retraction.
Therefore, it is necessary that the scientific community and the publishing groups reinforce the current editorial standards and particularly: (1) require raw data and/or open publication of the original data from any study; (2) require the study protocol and the statistical analyses prospective plan; (3) careful checking of authors' declarations and conflicts of interest; and (4) in-depth review of both study data/content and authorship.

**BLOCKCHAIN TECHNOLOGY TO SUPPORT THE DATA PROCESS IN SCIENTIFIC RESEARCH**

Until today, most systems used in research were based on traditional databases, that is, on centralised data stores without any built-in mechanism to ensure data immutability. Although these systems are very efficient to store large amounts of data, they have proven to be inadequate to prevent the problems mentioned above. Surgisphere itself developed a cloud-based traditional database of hospital records that was used for research.17 18

The facts evidently corroborate the idea that traditional data stores that do not guarantee the immutability of data are not sufficient to guarantee reliability in the data process, and it is probably necessary to evaluate new options.

To this end, blockchain has emerged as a technology that can guarantee the immutability, transparency and traceability of data even between two or more distant parties with no mutual trust.19 20 This technology seems particularly suitable for overcoming the problems described so far, having the capacity of guaranteeing quality and ethics in scientific research, without additional human supervision.

It should be highlighted that blockchain is a term that identifies several technologies that, combined together, result in a distributed, immutable and traceable ledger (the database).

One of the fundamental characteristics of these systems is that they are based on hashing algorithms to ensure data integrity. A hash is a string of letters and numbers produced by a specialised algorithm that converts a data input into a value of a fixed length that uniquely identifies the input. The algorithm can ensure that for each input a unique hash is generated, and that, given different inputs, two identical hashes cannot exist (conflict free). In blockchains hashes are used to concatenate the blocks of the blockchain, so that each block is uniquely identified and uniquely linked to the subsequent blocks, and also to uniquely identify pieces of data, for example, a document, in a secure and private way.

Public (permissionless) and private (permissioned) ledger solutions have been proposed. Public solutions, of which the most famous and widespread example is represented by the Bitcoin’s Blockchain, are open, do not have a ‘property’ or a third-party authorisation and are designed not to be controlled by any single entity. The purpose of the public is to allow everyone to contribute to updating data on the ledger and to have, as a participant, all immutable copies of all operations. That is, to have all the identical copies of everything that is recorded and approved on the ledger. However, this solution has some disadvantages, in particular that all transactions recorded on the ledger are visible to everyone, so confidentiality of information cannot be fully guaranteed. Additionally, public blockchains are based on consensus algorithms that are not suitable in contexts in which the number of users is very limited, such as in research studies.21

On the contrary, private blockchains rely on closed networks. Private blockchains are populated by actors who must rigorously share the same rules, and can easily be hosted on a single server, just as traditional databases. For example, Amazon Web Services currently offers an Amazon Managed Blockchain that makes it easy to create and manage scalable blockchain networks and distributed ledger technology, as also does Microsoft Azure, and others. This means that the transition towards blockchain-based solutions can take place gradually. Initially, systems can be structured as traditional cloud services, but with the advantage of relying on an immutable and fully traceable database. Then, once these systems have proven their merits, they can be extended to fully take advantage of the distributed nature of blockchain, by sharing the ledger with all the participants of the network.

In research applications, a private solution seems to represent a better alternative. It can be used for secure data collection, management and sharing. For instance, several decentralised data management solutions have been proposed, for example, to share electronic medical records between patients and providers. Choudhury et al developed a decentralised framework for consent management and secondary use of research data.22 Recently, a study on a blockchain-based software solution for clinical trials was conducted at Stanford University.23 The authors used raw data from a real completed clinical trial to simulate it onto a proof-of-concept distributed platform, testing its resilience to data tampering and providing a traceable and useful audit trail of trial data for regulators. Therefore, practical uses of blockchain solutions for enhancing the scientific research process’ reliability have already been suggested, and some adopted.21-23

Currently research databases are stored on the device of a single individual (or in a cloud system), of a research structure or of a healthcare system, where it is ‘fed’ by different, geographically distant subjects. In addition to the problem of data reliability highlighted by the cases of the *Lancet* and NEJM, this system also has the drawback that researchers often do not have access to the full database when they need to because they cannot access the device on which the full database is stored. By using blockchain this problem can be fully overcome.

Blockchain can allow complete traceability of the data (anonymised at the source), with creation/modification date, location and subject that originated the data, or possibly altered it. This can give a ‘license of reliability’ to the data, as long as from the early stages of the collection it is all managed through blockchain-based platforms. On this aspect, the editorial groups or the final stakeholders of figure 1 should incentivise researchers to adopt blockchain-based systems, right from the research protocol drafting, for example, by privileging for publication studies with ‘license of reliability’.

Moreover, the entire database can be stored on the device of each participant in the research project. At each datum or file change the database is continuously updated on all devices, so that all data are always synchronised. And since this mechanism can be based on an immutable ledger, everything is traceable and secured.

For all these reasons, we advocate that blockchains may have a place in research data collections. However, further studies about the on-the-ground deployment of such solutions are needed in order to test their feasibility and practicality in mitigating the risk of research misconduct. Moreover, given the early (prototype) stage of blockchains, their deployment must take place gradually, through a stepwise integration into the risk mitigation systems already used by the journals and described in the previous section (Strengthening existing journal policies).
CONCLUSIONS
Recent scandals highlight some serious concerns on the scientific integrity of today’s publication system. This situation, which is certainly not the first of its kind—and unlikely to be the last—confirms that ensuring data integrity is a professional and ethical obligation, aimed at providing reliable results to healthcare systems and regulatory authorities.

In other words, it is essential to implement a secure and reliable system, one that can be deployed within any research environment, that can ensure the traceability of the ‘core’ activities of medical scientific research. Such a system should mitigate the risk of frauds and misconduct and ensure that data are not altered, manipulated or falsified, or that—in case any of these actions are performed—it is possible to identify exactly who manipulated the data, by also tracing when and how.

The entire process of creation, collection and sharing of scientific research data, as well as the editorial one, should be reviewed. To do this, blockchain technology can be one implementable and scalable solution that can ensure data integrity in all phases of research, together with traceability and constant monitoring of data. However, their implementation should be integrated into the risk mitigation systems already used by scientific journals, which must necessarily be strengthened. This may help increase confidence and trust in the data and resulting evidence, both by the scientific community and the general public.

It is therefore essential to take advantage of recent events to rethink the scientific research system, starting in the biomedical field, by improving existing processes and adopting technologies aimed to reduce the risk and to increase the credibility of science itself.

Contributors All authors equally contributed to the study and to the manuscript. EB and DG conceived the idea of the study, reviewed the literature and wrote the paper. DG and GC conceived, wrote and reviewed the technical contents of the paper. MPF reviewed the manuscript for intellectual content.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement statement Not required

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

This article is made freely available for use in accordance with BMJ’s website terms and conditions for the duration of the covid-19 pandemic or until otherwise determined by BMJ. You may use, download and print the article for any lawful, non-commercial purpose (including text and data mining) provided that all copyright notices and trade marks are retained.

ORCID ID Davide Golinelli http://orcid.org/0000-0001-7331-9520

REFERENCES