CANCER GENOMICS & PROTEOMICS *20*: 211-221 (2023)
doi:10.21873/cgp.20376

Review

Biospecimen Digital Twins: Moving from a "High Quality" to a "Fit-for-Purpose" Concept in the Era of Omics Sciences

UMBERTO NANNI[1*], PATRIZIA FERRONI[2,3*], SILVIA RIONDINO[4], ANTONELLA SPILA[2,3], MARIA GIOVANNA VALENTE[2], GIROLAMO DEL MONTE[5], MARIO ROSELLI[4] and FIORELLA GUADAGNI[2,3]

[1]*Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy;*
[2]*InterInstitutional Multidisciplinary Biobank (BioBIM), IRCCS San Raffaele Roma, Research Centre, Rome, Italy;*
[3]*Department of Human Sciences and Quality of Life Promotion, San Raffaele Roma Open University, Rome, Italy;*
[4]*Department of Systems Medicine, Medical Oncology, University of Rome, Rome, Italy;*
[5]*Department of Palliative Care, San Raffaele Cassino, Clinical Center, Cassino, Italy*

Abstract. *The growing demand for personalized medicine we are currently witnessing has given rise to more in-depth research in the field of biomarker discovery and, thus, in biological banks that hold the ability to process, collect, store, and distribute "high-quality" biological specimens. However, the notion of "specimen quality" is subject to change with technological advancements. In this perspective, we propose that the notion of sample quality should shift from a broad definition of "high-quality" to a "fit-for-purpose" concept more suitable for precision medicine studies. Digital twins are a digital replica of real entities. These are largely adopted in any digitalized domain and are currently finding applications in biomedicine. The adoption of digital twins for biosamples, proposed in this paper, can provide prompt information about the whole lifecycle of the physical twin (i.e., the biosample) and substantially extend the possible matching criteria between the available samples and the researchers' and physicians' requests. This fine-tuning matching could greatly contribute to improving the "fit-for-purpose" quality, not only for studies based on current needs, but also to improve the identification of the best available samples in future situations, determined by the evolution of technologies and biosciences. Assuming and exploiting a data-science view in our biobank perspective, the more (accurate) data there are available, the more information can be extrapolated from them, the more opportunities there are for matching future, currently unknown, needs. This should be a mandatory principle that the 'time machines' called biobanks should follow.*

In recent decades, we have witnessed a substantial change in the approach to medicine from a global approach, considering the various clinical entities at the same level, to a growing demand for personalized medicine. Oncology in particular, is one of the fields most demanding for personalized/precision medicine, as highlighted in the 2016 recommendations of the Blue-Ribbon Panel of the Cancer Moonshot initiative (1).

This new vision of medicine has given rise to more in-depth research in the field of the so-called omics (proteomics, peptidomics, lipidomics, metabolomics, transcriptomics, and the more recent radiomics and pathomics), enabling translational studies geared toward the search for new molecular predictive/prognostic biomarkers, ultimately fostering the field of biomarker discovery. The integration of computer data derived from biomarker discovery studies has, in turn, helped to define new algorithms for estimating risks to be applied in a personalized clinical approach that would ultimately enable the prompt application of optimized treatment protocols for each individual patient and a more rational use of drugs (2).

One lesson in this direction comes from the clinical application of personalized medicine approaches in the management of cancer patients, which represents a key area of care worldwide. Cancer incidence and prevalence, in fact,

*These Authors contributed equally to this study.

Correspondence to: Fiorella Guadagni, IRCCS San Raffaele Roma, Research Centre, Via di Val Cannuta 247, 00166 Rome, Italy. Tel: +39 06-52253733, e-mail: fiorella.guadagni@sanraffaele.it

Key Words: Biobank, biospecimen science, sample quality, personalized medicine, digital twins, review.

are globally rising, boosted by an increasing aging population, negative lifestyle choices and exposure to environmental factors, among others. Much has been done in terms of prevention strategies and early diagnosis and screening programs, yet cancer still represents a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, according to the latest WHO estimates. To counteract this trend, new therapeutic approaches have been introduced, which have resulted in an increase in survival and a reduction in mortality from cancer, in the face of an increased burden of care-related costs. Conventional therapeutic approaches based on the choice of chemotherapy relying on histopathological evaluation of the tumour are, indeed, out-dated, in an era of molecular phenotype measurements and characterization (through genomic and/or proteomic approaches) and precision oncology (3). Personalized treatment for cancer is today limited to a rather small number of drugs towards molecular targets. However, significant additions to the currently available armamentarium are rapidly developing from the execution of biomarker discovery studies based on high-throughput technologies. These are aimed at identifying new molecules as determinants of outcome response rates, ultimately ensuring survival and safety, predicting patients' failure to respond to drug treatment, or identifying and minimizing the occurrence of side-effects.

All the above have allowed the design of projects involving many biological samples from a great number of individuals, well stratified for a given condition being either affected, vectors, or predisposed to genetic or environmental diseases, or who display variable responses to treatments and that can be compared with control groups matched for different characteristics. As a result, the amount of stored material has significantly increased, giving life to the new concept of a biological sample with its associated clinical data, representing a fundamental step forward towards the optimal design of any research project.

Relevance of Biobanking

In this context, given the potential applications of biomarker discovery studies to the pharmaceutical, biotechnology and bioinformatics field, it is logical to understand the great interest of biomedical research in biological banks. In 1999, the Swedish Medical Research Council first provided a definition of a research biobank as a "*collection of human tissue samples, the origin of which can always be traced, stored for a defined or indefinite period for specific study projects*" (4). Initially, these collections were small, mainly university-based archives, created and developed to meet specific local projects' needs. Since then, the field of biobanks has grown and enabled improved scientific research to such an extent that, already in 2009, the Time magazine enlisted research biobanks among the 10 discoveries that might have changed the world (5). Hence, the use of the term "biobanking" envisioned as the acquisition, processing, and maintenance of human samples for research began to play a key role, progressing enormously to become a science in its own right. It therefore became essential that the small archives be transformed into institutionally organized collections.

Biological banks, now defined by the Oviedo Convention as "operational units that provide a service for the preservation and management of biological material and associated clinical data in accordance with good laboratory practice, privacy law and ethics guidelines", constitute a fundamental resource, even many years after sample collection (6).

These facilities, in part thanks to the implementation of Standard Operating Procedures (SOPs), the harmonization of available information, along with the development of data integration processes, are particularly well suited for the growing demand for biological specimens to include in research protocols (7).

Ensuring Sample Quality, a Matter of Extreme Importance.

The Quest for Biosample Quality

The availability of large numbers of biological samples demands that they be homogeneous in terms of pathology, clinical and collection characteristics, as well as storage procedures. This requirement has assumed such a crucial role in the field of biomedical research that a survey, conducted among a substantial number of investigators, showed that many research activities are severely hampered by the heterogeneity and the heterogeneous quality of the human samples used (7, 8). Currently, the heterogeneity in the quality of collected biomaterials is such that it may even hinder the development of more effective therapies and diagnostic tools (9, 10). Given that also biobanks are heterogeneous in their design, content and use, a high degree of harmonization and standardization in the various biobanking activities should be achieved.

The need to solve this problem has prompted several international scientific societies to produce a set of high-profile, well-defined "best practice" guidelines appropriate to the type of biological material being collected. One of the first published documents is by the International Society for Biological and Environmental Repositories (ISBER). Since its first publication in 2005 and in subsequent revisions – the most recent that of 2018 – the ISBER document has provided best practices for technical issues that have arisen during the evolution of biorepositories and biobanks (11, 12). Other Institutions such the US National Cancer Institute (NCI) firstly published the "NCI Best Practices for Biospecimen Resources" on 2007 (13) and posted on the NCI Biorepositories and

Biospecimen Research Branch (BBRB) website the 2016 revision, intended to provide more detailed recommendations related to biospecimen and data quality.

Several initiatives for the harmonization and standardization of biobanks have been also carried out at the European level by the Organisation for Economic Co-operation and Development (OECD) (14) and the European research infrastructure for biobanking (BBMRI) (15). All these initiatives resulted in a demand for an industrial standard, requiring a formal qualification of the automated storage system before it is put into use, as well as frequent re-evaluation of its performance with the ultimate goal of ensuring sample quality and reproducibility necessary to facilitate the advancement of translational research. ISO Standard 20387:2018 was developed with the goal of providing biobanks with a line to follow to enable them to collect biological material and associated data of adequate quality for research and development (16). More recently, due to a growing attention to the crucial role of biobanks, ISO/TR 22758:2020 has been issued, detailing the requirements for the competence, impartiality, and consistent operation of biobanks, which is intended to be a supplement to, rather than a substitute for, ISO 20387 (17).

Indeed, a biobank's ability to process, collect, store, and distribute "high-quality" biological specimens is based on the meticulous application of SOPs and evidence-based best practices. It is now well known that the "quality" of a biological sample used for biomarker discovery studies is a multidimensional concept, depending on many factors, foremost among them in terms of impact, the pre-analytical stage. Lack of information of some processes in this phase can significantly affect the research results themselves. This is particularly relevant in the field of omics where technologies are very sophisticated and constantly evolving. It is, therefore, necessary for biobanks to provide the researcher with the traceability of pathways and processes that the sample has undergone over time. With this in mind, the demand for instrumentation comes, that will allow complete traceability of the life of the biological sample – whether liquid- or tissue-derived.

The Development of a Standard PREanalytical Code

Several studies have shown how small variations in the pre-analytical stage of the biological sample can have a huge impact on the results obtained in translational research protocols, particularly when using the new omics technologies, whose tests are generally characterized by a high sensitivity (18). Therefore, the use of automation and standardization approaches that can ensure complete traceability of the biological sample lifecycle plays a key role in the value chain of translational research and represent effective tools to be more rapidly incorporated for the significant developments in the field of precision medicine (19). The NCI, in its 2016 publication of best practices for biospecimen resources, provided useful guidance to minimize the effects of manipulation on the integrity of the sample during acquisition and preparation phases that might significantly impact on the reliability of analytical results. The document reiterated the need for more detailed information on samples used for research activities, with particular regard to monitoring the lifecycle of each biological sample in order to trace carefully intentional and unintentional pre-analytical variations, resulting from sample collection, preparation, or storage (13). These guidelines all converged on a few main conclusions: standardization, harmonization (SOPs) and process control, especially through automation of the most critical areas for biobanking and more detailed information to ensure comparability of sample features across networks. The goal is to provide the researcher with a set of samples that are "equivalent" for the intended use, that is, the preanalytical variables are either identical or their differences are irrelevant for the planned tests. In this light, it is important to note that even the use of different SOPs might impair the comparison of results among studies; therefore, the more precise the recording of the preanalytical parameters of different types of biosamples, the more accurate will be the extraction of valid information when required for clinical or research purposes.

A solution to simplify this need was first proposed by the ISBER Biospecimen Science Working Group in 2009 with the presentation of a "Standard PREanalytical Code" (SPREC) which is now used internationally to precisely define these characteristics considered indispensable for the subsequent analytical phase (20). The SPREC of a sample is a coding of the sample preanalytical history, which can be displayed, for example, as a barcode or a QR-code. In the resulting code, the most relevant preanalytical variables (currently acknowledged) of liquid and solid samples and their sample derivatives are reported, which – together with the accompanying sample and biomedical data – constitute the scientific value. Briefly, the SPREC consists of seven numerical elements, each corresponding to a preanalytical variable, and a sequence of letters (different for fluid and tissue samples), that can provide the detailed information about the preanalytical collection, preparation, and storage procedures to which the individual sample was subjected (20). Because of its versatility and easy implementation, SPREC is continuously updated (21) and adapted to the requirements of new technologies (22), thus allowing biobanks to include new matrices or new preanalytical options.

IT-based tracking of storage conditions. In addition, the extent to which biological samples are adversely affected by processing conditions (out-of-range temperatures, multiple freeze-thaw cycles) should not be forgotten. Inadequate processing, handling, prolonged interval between processing

steps, and storage can severely affect the quality of the biological material (23, 24). Based on these observations, the Multidisciplinary Inter-Institutional Biobank of the IRCCS San Raffaele Roma, Italy (BioBIM®) developed a pilot study, based on IT (Information Technology), for tracking the steps following pre-analytical processing, which is usually not tracked by automated tools, as it is manually managed. Detailed data were collected regarding the entire lifecycle of stored samples using radio frequency identification (RFID) technology (25). The first was to analyse the processing chain of blood samples that is operational at the BioBIM. Research focused on the problem of traceability of the steps following automated preanalytical processing: digital records of the time when an event occurred that could affect the biological quality of the samples were collected using RFID tags and readouts. In a pilot study of 2011, the storage conditions were traced between the completion of the pre-analytical phase and the analytical phase, *i.e.*, a time interval of the lifecycle which is not usually tracked by automated tools because it typically includes manual handling. By adopting RFID devices, not only the possible critical timelines were identified, but the functionality of the system over time was also demonstrated. Indeed, the procedure performed at 1-, 3-, 6- and 12-month storage showed that RFID-labelled samples cryopreserved at -80°C could still be successfully read (25). Tests have been continuously performed over time showing that RFID are still readable after 10 years (Fiorella Guadagni, personal communication, data recorded in June 2022).

Specific Challenges or Gaps

Security issues and data protection compliance. The technological advancement in biomedical science achieved with the development of omics technologies and infrastructure such as biobanks, has made the issue of data security of utmost importance. Within the context of biobanking therefore, it becomes mandatory to identify the stakeholders involved and to define how the data will be processed, the measures put in place for protecting the data and the technical and organizational measures that enable end-to-end data protection, from collection to use, to sharing of results. In addition, there is increasing recognition of the need to ensure the use and reuse of data, which poses additional issues in terms of FAIR Principles and Ethical Aspects that need to be considered when using biosample-associated data.

All these require that data sharing be done with consideration of all aspects related to patient privacy and consent, as well as the proper use of the data that must be regulated according to National and International regulations. We must keep in mind that, until now, informed consent forms generally did not contemplate the possibility that data could be publicly shared with others and, therefore, patients did not explicitly choose to share data publicly. Furthermore, failure to include an opt-in statement for public data sharing in the consent form does not constitute tacit approval for public data sharing. In this respect, we must agree with the concerns raised in Michael C. Gibson's viewpoint that "*data could be used in a way that was never intended by patients or researchers and could result in unforeseen damages*" (26).

The data transformation from input sources to the outputs produced presents many challenges in terms of ensuring, at each stage of the process, that data are available exclusively to pre-authorized participants, and that data quality is constantly monitored. This generally implies that the risk to confidentiality and quality of the data is minimized. More specifically, for software and systems deployed in the biobank, it will be necessary to define data security requirements in terms of logistical security of data communications from sources (*e.g.*, biobanks) to any system involved and its individual components. Moreover, mapping of authorized accesses, logic monitoring of functional and technical data to guarantee non-alteration and logic tracking of the data transformation should be taken into consideration. Finally, procedures for analysing the adequacy of processing and scalability of data or possible security breaches must be defined.

Security should be granted during the entire lifecycle of the biosamples/data, while visibility and transparency of biobanking processing should be ensured, in order to render data protection verifiable. Finally, a Data Privacy Impact Analysis (DPIA) should be performed, taking into account both the known risks and the technical and organizational measures that must be adopted to mitigate those risks.

The need for a federated protocol for searching samples. Despite the amount of financial resources, multidisciplinary research activities and organizational efforts devoted to the creation and enhancement of biobanks worldwide, and despite the SOPs adopted and the dedicated networking organizations created, the currently available way of searching for biosamples still requires an expensive investment of manual time in the face of several possibly well-digitized biorepository websites. More explicitly, a researcher who needs collecting samples, usually must deal with a number of information sources, such as the websites of the available biobanks. Then, each biobank presents a catalogue of the available samples and possible services or – at best – it has its own search engine, based on the criteria deriving from the internal Biobank Information System. This second approach is preferred by the biobank itself, since it does not necessarily disclose explicit information that is not intended to become public.

Each biorepository, of course, has its own information system, and information regarding samples is usually spread over a number of tables and/or subsystems. The difficulties in finding specimens for potential users may have, consequently, a low usage of the valuable resources stored in biobanks, and ultimately hinder their sustainability.

A possible approach to reduce the burden of sample retrieval could be based on a federated approach. In a federated information system, there is a network of servers, each owning private information open to queries, without a central database. A user can formulate a single "query" (by a suitable interface and template) that is accepted and interpreted by all the servers participating in the federated network. The final answer to the user comes after collecting the answers from all servers. Of course, this model is difficult to implement when the network is heterogeneous, *i.e.*, each node – in this case each biobank – has its own information system. A full integration of the information systems is out of the question. However, a federated architecture is sufficient, *i.e.*, a network where each participant has the ownership of its own information and decides what services are to be provided to external users with a common protocol. Of course, the starting point for a common protocol is a set of shared and well-understood concepts: biosample digital twins can be a solid base to make such scenario operational.

Future Perspectives: Empowering Biosamples With Digital Twins

The digital twin of a specimen. In several areas, the notion of a "digital twin" of a real entity has been established as a solid methodological approach to manage the features of a real entity in many possible ways, in order "to mirror the life of its (physical) twin" (27). Introduced by John Vickers of NASA in 2010 (24), many improvements derive from this approach, in terms of feature conceptualization, data integration from heterogeneous sources, integration with AI techniques, and a possible strong movement to support the digitalization of domains and industries.

The seminal contribution (28) "introduces the concept of a digital twin as a virtual representation of what has been produced. Compare a digital twin to its engineering design to better understand what was produced versus what was designed, tightening the loop between design and execution". Formulated originally in the realm of manufacturing, and focusing on the importance of the digital twin to trace the whole lifecycle of the physical twin (*i.e.*, the corresponding physical object), Grieves highlights the openness of this concept to: (a) the evolution of complexity of our needs, and (b) our capacity to collect information in the physical world: "the amount and quality of information about the virtual and physical product have progressed rapidly [...]. The issue is that the two-way connection between real and virtual space has been lagging behind." Years later, this gap has now been filled by what has been named operational technology (OT), concerning the exchange of information between the real world and information systems.

Not surprisingly, proposals of medical digital twins are spreading in literature. Examples include: modelling the human immune system (29), human heart (30), food products for diabetics (31), multiple sclerosis (32), drug discovery (33), or associated with patients see, for example, the "Personal Digital Twin" (34), engineered cells (35), and various other concepts. Focusing on biorepositories, to the best of our knowledge, there is no evidence in literature of the adoption of digital twins associated with biosamples. Nonetheless, we strongly believe that they may help creating a digital ecosystem supporting research efforts in network medicine. A graphical representation of the relevant properties of digital twins in biobanking is depicted in Figure 1. As shown, a digital twin of a specimen should consist of the collection of information associated with the physical twin (*i.e.*, the real specimen) – including its origin, consistency, preparation, exact location, physical constraints – as well as the documentation of the patient's consent to data usage (digital documents, in an original or digitally signed copy). Information on the specimen lifecycle (including the storage relevant events and documents) and any results (data, images, textual annotations, etc.) obtained from any test carried out on the sample should also enrich the information set, together with all the clinical information obtained in full compliance with existing regulations and the patient consent.

We point out that the digital twin of a sample generally evolves in time, reflecting the lifecycle of the physical sample, is usable by itself and has an unlimited durability, even beyond the lifespan of its physical twin that may degrade, or cease to exist. Accordingly, the value of a (physical) specimen depends heavily on its digital twin and cannot be used without (at least a portion of) its digital twin, whereas the digital twin is usable by itself.

A digital twin should be easily built from a local computation by any information system holding a full traceability of the physical twin, which brings our attention on the crucial informational content – which is a substantial part of a biobank – and provides a practical way to deal with the problems related to proper data management, complying norms, internal procedures, and patients' will. As an example, when a biobank provides a researcher with a (physical) specimen, this can be accompanied by a portion of the original digital twin, depending on the specific use, the contracted agreement, complying the norms in effects, the biobank regulations, and the patients' consent for that specific use of the specimen.

A federated architecture for exploiting biosamples' digital twins. Usually, biobanks have their main motivating target in making their collections searchable by possible users, for translational research purposes. On the other side, while pursuing their targets, researchers have a corresponding interest in searching over the largest possible biobanks' collections. Digital twins are a step toward a convergence of

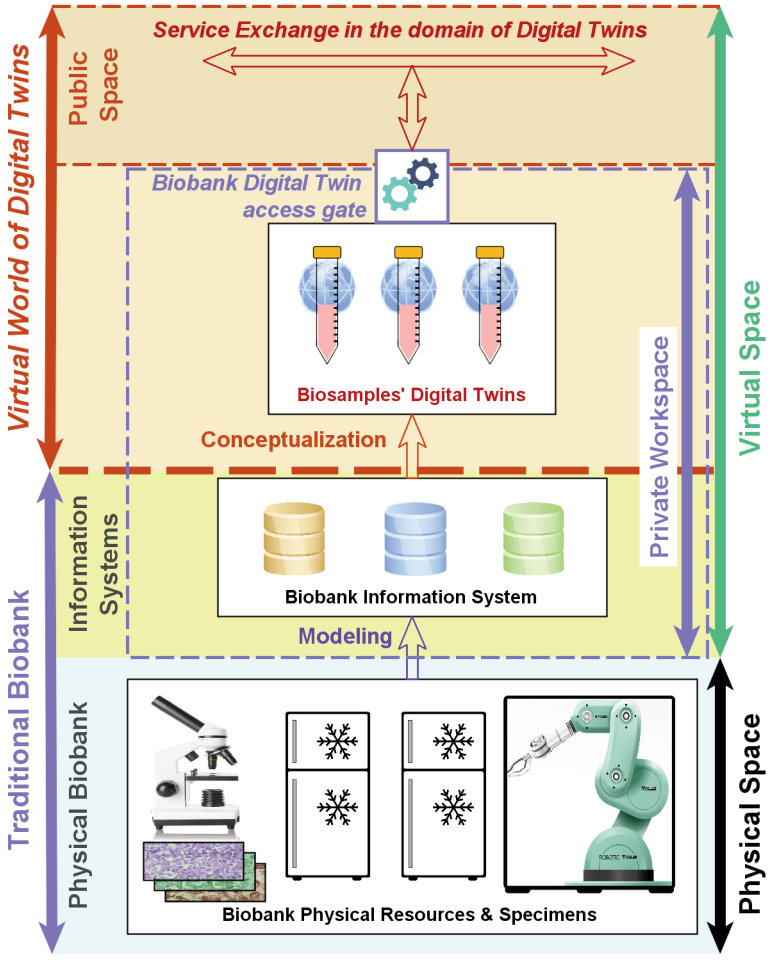

Figure 1. *Digital twins of specimens: a representation of the relevant properties of samples. The digital representation of specimens incorporates documentation about the specimens' origin, the patient consent, and the possible use, pre-processing and storage conditions, as well as any other detail characterizing the sample. Each biobank can decide the features of its own biosamples that are to be exposed to the public, or to specific users. The specimens' digital twins can be matched with search criteria and can enforce security and privacy regulations; they may encompass existing standards [e.g., Standard PREanalytical Code (SPREC) (20)] and encourage further standardization. The specimens' digital twins are assumed to populate, together with other entities (pathologies, drugs etc.) a virtual world of digital twins – a digital ecosystem supporting research efforts in network medicine.*

the two goals. The adoption of digital twins could be a crucial step to encourage the creation of a digital ecosystem of biobank operators and users, in which biobanks create digital twins containing both purely internal interest information (*e.g.*, physical location of the physical twins, references to internal processing and layout etc.), and public interest information defining the user's view, including the objective features that might match the user's needs, conceived for all possible users, possibly including unusual and/or future needs.

The materialization of digital twins, and the existence of a user's view, based on conceptual models defined by real biomedical needs and ignoring all internal features of the physical infrastructures (including the internal biobank information system), enables rapid processing, within the virtual world of digital twins, according to the user's requests. Further evolutions of such needs –intrinsic to the advancement of science and technology – may find details already available in digital twins in biobanks with a richer and more detailed sample lifecycle data collection, or push biobanks to improve their internal processes and/or extend the information collected in the digital twins with the required details.

The digital twins in biomedical sciences can provide much more than a mere collection of already available information: this notion might stimulate the evolution toward a common mind-set in complex domains, where an explicit representation of relevant concepts and scenarios stimulates sharable visions and insights, while it provides promptly

processable data. In the case of biosample digital twins, we will show as these, by means of adequate protocols, can be a solid base for supporting identity, accountability, and compliance to regulations of these entities characterized by highly sensible data. A simple but comprehensive overview of biomedical digital twins was recently provided by a seminal presentation by Eric A. Stahlberg, the director of NCI Center for Cancer Research Bioinformatics Core (36).

The first steps toward a standardization of information associated to biobank samples have attracted growing interest, as commented above. Following this same trend, a possible future scenario and information flows result from the adoption of digital twins supported by a federated protocol. An interesting example of implemented federated architecture for sharing biosamples and various information services within a community of cooperating institutions has been proposed, for example, in (37). Here we propose the main features of a solution based on digital twins within a federated architecture that would implement an open protocol with two roles for participants: biobanks, holding the biosamples and related information, and end users, with research or medical objectives. All share one common goal: to enable users to find the desired samples, accompanied by the necessary data, in some biobank in the network. In principle, all nodes in the network – both end users and biobanks – can be independent entities. An agreement would be necessary only if, and when, a biosample or data transfer actually has to take place. Of course, more complex and close forms of cooperation are possible, but devising open and light interaction outside of predefined agreements is one possible way to enhance biosample and data retrieval on a global scale, something that seems far from being a reality today.

A basic architecture, connected to the various Biobank Information Systems is reported in Figure 2. As shown, each biobank willing to enter the Digital Twin ecosystem characterizes its specimens by creating the corresponding Digital Twins (Figure 2, node 1), which can still be retained in a private workspace; this information could only be accessed in terms of responding to sample retrieval queries. End users can characterize their needs by providing the required sample features and quantity by means of one or more (partially specified) “Query Sample”, possibly characterized by alternatives or ranges (Figure 2, node 2), which will be handled as distributed queries in the federated architecture and delivered to all the compliant and connected Biobank Digital Twin access gates (Figure 2, node 3). Each biobank (or any organization in charge to perform this task on behalf of the biobank) matches – in a private space (Figure 2, node 4) – each Query Sample with the current collection of digital twins, looking for (possibly partial) matches. The answers from all the biobanks are collected and returned to the user, with all the required accessory information (Figure 2, node 5). In principle, besides basic matching, each biobank might use AI techniques to understand whether (and how) the desired samples can be prepared by suitable processing of existing samples in the biobank.

If the agreement is completed, and the physical samples are delivered to the researcher or physician, then the corresponding relevant and meaningful portion of the digital twins are sent to the requester. As a possible element of the agreement, the results obtained by the requester (*i.e.*, the shared portion of the Digital Twin updated with the result of testing and data processing) could be returned to the originating biobank, thereby enriching the original digital twins of the biobank.

Impact of digital twins on the protection and security of data. Concerning security, nothing comes as a straightforward consequence of the adoption of digital twins, unless norms, and/or standards will regulate these entities. This representation is a relevant step towards the adoption and the enforcement of uniform politics for complying security and privacy norms. First of all, this is a first step towards the definition of standard representations: these can be defined at a conceptual level (*i.e.*, it is not just a technical format, such as EDI), hence can be made clean and fully understandable to experts in biomedical areas, who can discuss and agree on the substantial contents. A digital twin, beside any information concerning the physical twin (the sample itself, its lifecycle and current state, results of previous tests, and so on), may include, as an example, any sort of document, such as: (a) the original patient's informed consent (with possible updates), with legal digital signature, or a digital legalized copy of it; (b) any bilateral agreement concerning the provision of (a portion of) a sample to any authorized entity (this may concern the physical sample and/or its digital twin), so that for any physical or digital transfer the compliance with the patient's consent can be verified; (c) the responsibility of maintaining this documentation can be accounted to the original holder of the sample and made consultable to authorities in charge; (d) any organization holding a physical sample and or a (portion of) a physical twin, either must be the original sample holder, or must provide a link to the "master digital twin", providing the evidence of the regular possession. We point out that, in this scenario, any novel definition and/or update of regulations, based on clear terms that any expert in biomedical areas will understand, can rely upon such robust arrangement of accountability; any organization holding samples will share with any specific peer or making public only the information that it is intended to be shared, and - on the authorities' request - must be able to prove that such use of information is compliant to the original (or renewed) patient's consent. Indeed, the adoption of digital twins (as digital entities) requires the establishment of procedures ensuring data confidentiality and security compliance. More

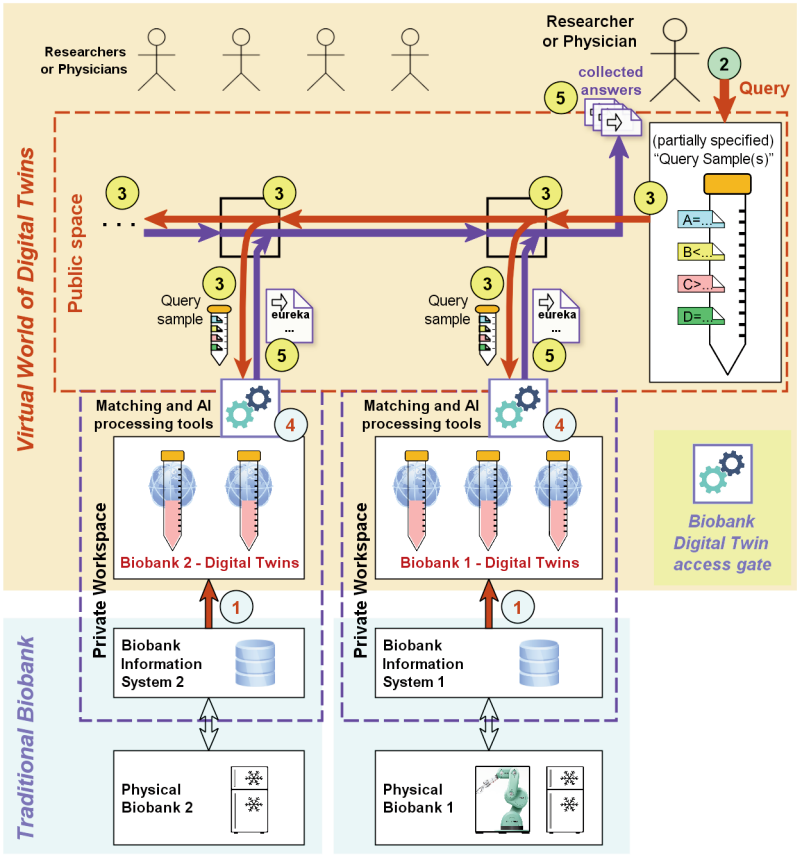

Figure 2. *Querying the ecosystem of digital twins supported by a federated protocol. The ecosystem envisages that each biobank can build – in its own private workspace – the Digital Twins of the (physical) available specimens it holds (1). On the other hand, end users can formulate a user query [e.g., one or more query sample(s)], based on actual needs, and ignore the internal details of each biobank collection (2). The query is, then, sent around the federated network and delivered to each available biobank [accessible through a Digital Twin access gate (3)], which can process it in its own private workspace, finding the matching digital samples, or determining the possibility of providing (some of) the desired items by suitable processing of existing specimens (4). The answers from all available and interested biobanks are finally collected and returned to the users (5).*

specifically, any institution that is an original holder of specimens providing an external availability of the individual specimens and/or related data, would be in charge for the management of data related to the biosamples.

A characterization of a confidentiality and security management system for biosamples is far beyond the scope of this paper, but we can provide some evidence of the boost provided for this purpose by the biosamples digital twins. We list some of the main requirements that could establish a base for enforcing security compliance – focusing on features that biosample digital twins make viable. This list is conceived to envision a possible scenario, suitable to encompass the evolution of norms and recommendations of the proper authorities, as well as of the relevant standards that will be defined in this area. First of all, when a biosample is prepared for storage in a biobank, a corresponding "Master Digital Twin" must be created, which incorporates all the relevant information, including the digitally signed patient's consent, or an authenticated copy of the original document. Clearly, any personal information must be stored only under strong encryption. Secondly, each access to a physical biosample or to its digital twin must take place within one of the allowed procedures and must be compliant with the pertinent regulations. In particular, any transfer can take place only in compliance with a material transfer agreement (or data transfer agreement) between the sending and the receiving institutions. Finally, any exchange of specimens and/or related data must be safely traced (*e.g.*, relying upon blockchain technology); as a consequence, any human specimen, in any biobank or research laboratory, either has an internal origin and has never been moved, or has been imported through a traced transaction; in both cases, on request by the proper authority, the origin of the sample can be assessed, as well as any pertinent Transfer Agreement

with the receiver commitments. Thus, security, visibility and transparency can be granted during the entire lifecycle of the biosample and/or the attached data.

The above requirements imply not only technicalities that biobanks must deal with, but also the role of the authorities in charge of their regulatory framework – that should be settled and updated at the international level, and with reasonable frequency. The evolution of data processing techniques may create subtle distinctions among the possible forms of data exploitation (*e.g.*, federated learning). In order to avoid creative interpretations of obsolete rules, both regulations and patient's consent formulation must evolve accordingly. This is a key activity to maintain, in a secure environment, a proper equilibrium between research needs and patients' rights.

Moving from a "High Quality" to a "Fit-for-Purpose" Concept

In recent years, the approach to medicine has substantially changed, under the growing demand of providing precision medicine. Oncology, in particular, is one of the fields most demanding for precision medicine, as recommended by Members of the Blue-Ribbon Panel of the Cancer Moonshot initiative that launched, in September 2020, the U.S. NCI Cancer Moonshot Biobank (1) with "*the aim of accelerating cancer research by creating a resource of well-annotated, longitudinal cancer biospecimens from patients receiving standard-of-care therapy [...] to better understand the mechanisms and systems biology of drug resistance and sensitivity*".

The increasing demand for new biomarker discovery in the field of personalized medicine requires robust datasets that also rely on the availability of a large number of standardized specimens to the extent that many research activities are seriously invalidated by the heterogeneous "quality" of the samples used, which is often significant even within a given biobank, but certainly may be more important among different biobanks. Therefore, it becomes essential to gather information about a specimen and its processing that may enable the right options to be chosen for the right specimens and, vice versa, the right specimens for the target study. Indeed, the suitability of samples for a given research also depends on the type of test to be performed.

A further fundamental issue that a biobank must address is the durability of the stored specimens. Usability of samples for long periods, on the order of years or even decades, means dealing with technological advances in omics and biology-related sciences. What are the requirements that define a "good" specimen usable in laboratory testing several years from now? In this context, the best service we can provide to future physicians and researchers is to deliver samples with detailed documentation of their lifecycle, possibly well beyond current needs. The lifecycle affects the future "quality" of a specimen in ways we do not fully know– and never will – because the notion of "specimen quality" is subject to change with technological advances. Therefore, gathering more details about a sample's lifecycle means extending its possible usability in the future and hence, ultimately, its scientific value. It is therefore of utmost importance to change our perspectives on the definition of quality in the biospecimen science field, evolving from the concept of absolute "high-quality" sample to the new conception of "fit-for-purpose quality", where sample suitability is defined by the appropriateness for a given study/technology.

The notion of the digital twin focuses on the data that characterize the samples. This paves the way for a common representation, *i.e.*, a simple-to-be-adopted de facto standard and later – hopefully – for solid standards and regulations that will simplify a controlled data transfer among scientific institutions, with full compliance to privacy norms and patients' consent. As clarified in previous sections, this does not imply a common implementation of biobank information systems (BIS), but a shared user-oriented information model, independent of each biobank's information systems.

The possibility of exploitation of a specimen is ensured by the richness and completeness of its digital twin, and the identifiability of a specimen as "fit-for-purpose" is based on the matching performed at the conceptual level, in the virtual world of digital twins – regardless of the internal feature of the BIS that usually disseminates information relative to specimens (and related entities) in a series of subsystems. In real-world situations, the richer the "potential" documentation of a specimen, the more fragmented and sparser its documentation is. In practice, all this information is usually hidden and not promptly available to potential users, hindering potential uses of the specimen itself and preventing the selection of the "best" specimens that would optimize both the pursuit of the potential user's goal, the use of the biobank's resources and, ultimately, its sustainability.

A crucial point is that the digital twin approach does not imply sharing biobank insights. On request, this approach makes it easy to compare (external) user needs with the actual resources of the biobank (a comparison that takes place within a private workspace, where digital twins are stored) and select the portion of the digital twin to be shared with potential and end users of the physical samples.

Conclusion

In conclusion, we strongly believe that the concept of sample "quality" should shift from a broad definition of "high-quality" to a "fit-for-purpose" concept, more suitable for precision medicine studies. In this context, the adoption of digital twins associated with biosamples will encourage the creation of a digital biobank ecosystem – that is, a self-

organized community of users and operators – in which there is a clear separation between the internal organization of biobanks, including the Biobank Management System, and the AI-based exploitation technology, which operates in the domain of Digital Twins and focuses exclusively on users. A federated approach to sample retrieval will naturally let biobanks have full ownership of their data and possibly support non-disclosure policies for their internal resources. In this digital ecosystem users can focus on the features of specimens using their own vision, defined at a conceptual level, independently of the internal characteristics of the biobanks and their internal organization. On the other hand, biobanks will have the opportunity to increase the utilization of their resources and, ultimately, their sustainability. Finally, the need to provide services to advanced user needs should prompt biobanks to adopt a more accurate sample lifecycle tracking by pushing Operational Technology, which provides more cost-effective solutions for monitoring.

Better interaction between biobank operators and end users, together with sustainability motivations, will likely help promote the concept of "fit-for-purpose" quality not only for studies based on current know-hows, but also to identify appropriate samples for emerging technologies.

Conflicts of Interest

The Authors declare no conflicts of interest.

Authors' Contributions

U.N., P.F. and F.G. conceived and drafted the manuscript; S.R., A.S., M.G.V., and G.D.M. reviewed the manuscript; M.R. critically revised the manuscript for important intellectual content; F.G. provided funding for the study. All Authors have read and agreed to the published version of the manuscript.

Acknowledgements

The Authors express their deep gratitude to all patients and their families for providing the opportunity to conduct the Biobank project. Special thanks are due to the Rotary Club Latina San Marco for having been awarded the Biobank Project. Finally, authors wish to thank Luigi Narducci and Danilo Sayed Ibrahim for their excellent technical assistance in biobanking activities. This research was partially funded by the Italian Ministry of Health (Ricerca Corrente) and by the European Project Horizon 2020 SC1-BHC-02-2019 (REVERT, GA n. 848098).

References

1 Cancer Moonshot. Available at: https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative [Last accessed on December 21, 2022]

2 Drucker E and Krapfenbauer K: Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J *4(1)*: 7, 2013. PMID: 23442211. DOI: 10.1186/1878-5085-4-7

3 Gambardella V, Tarazona N, Cejalvo JM, Lombardi P, Huerta M, Roselló S, Fleitas T, Roda D and Cervantes A: Personalized medicine: Recent progress in cancer therapy. Cancers (Basel) *12(4)*: 1009, 2020. PMID: 32325878. DOI: 10.3390/cancers12041009

4 Swedish standards set for use of genetic 'biobanks'. Prof Ethics Rep *12(3)*: 3, 1999. PMID: 15584149.

5 Park A: Biobanks - 10 ideas changing the world right now - TIME magazine 2009. Available at: http://content.time.com/time/specials/packages/article/0,28804,1884779_1884782_1884766,00.html [Last accessed on December 21, 2022]

6 Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine. J Med Philos *25(2)*: 259-266, 2000. PMID: 10833140. DOI: 10.1076/0360-5310(200004)25:2;1-O;FT259

7 Betsou F, Luzergues A, Carter A, Geary P, Riegman P, Clark B, Morente M, Vaught J, Dhirr R and Druez-Vérité C: Towards norms for accreditation of biobanks for human health and medical research: Compilation of existing guidelines into an ISO certification/ accreditation norm-compatible format. Qual Ass J *11*: 221-294, 2007. DOI: 10.1002/qaj.425

8 Day JG, Iorenz M, Wilding TA, Friedl T, Harding K, Pröschold T, Brennan D, Müller J, Santos LM, Santos MF, Osório HC, Amaral R, Lukesova A, Hrouzek P, Lukes M, Elster J, Lukavsky J, Probert I, Ryan MJ and Benson EE: The use of physical and virtual infrastructures for the validation of algal cryopreservation methods in international culture collections. Cryo Letters *28(5)*: 359-376, 2007. PMID: 18075705.

9 Moore HM, Compton CC, Lim MD, Vaught J, Christiansen KN and Alper J: 2009 Biospecimen research network symposium: advancing cancer research through biospecimen science. Cancer Res *69(17)*: 6770-6772, 2009. PMID: 19706749. DOI: 10.1158/0008-5472.CAN-09-1795

10 Betsou F, Barnes R, Burke T, Coppola D, Desouza Y, Eliason J, Glazer B, Horsfall D, Kleeberger C, Lehmann S, Prasad A, Skubitz A, Somiari S, Gunter E and [International Society for Biological and Environmental Repositories (ISBER) Working Group on Biospecimen Science]: Human biospecimen research: experimental protocol and quality control tools. Cancer Epidemiol Biomarkers Prev *18(4)*: 1017-1025, 2009. PMID: 19336543. DOI: 10.1158/1055-9965.EPI-08-1231

11 International Society for Biological and Environmental Repositories (ISBER). Available at: https://www.isber.org/ [Last accessed on December 21, 2022]

12 Campbell LD, Astrin JJ, DeSouza Y, Giri J, Patel AA, Rawley-Payne M, Rush A and Sieffert N: The 2018 revision of the ISBER best practices: Summary of changes and the editorial team's development process. Biopreserv Biobank *16(1)*: 3-6, 2018. PMID: 29393664. DOI: 10.1089/bio.2018.0001

13 NCI Best Practices for Biospecimen Resources. Available at: https://biospecimens.cancer.gov/bestpractices/2016-NCIBestPractices.pdf [Last accessed on December 21, 2022]

14 Organisation for Economic Co-operation and Development (OECD). Available at: https://www.oecd.org/sti/emerging-tech/44054609.pdf [Last accessed on December 21, 2022]

15 European research infrastructure for biobanking (BBMRI). Available at: https://www.bbmri-eric.eu [Last accessed on December 21, 2022]

16 ISO(2018) ISO 20387:2018 International Standard: Biotechnology - Biobanking - General requirements for biobanking, First Edit. Available at: https://www.iso.org/standard/67888.html [Last accessed on December 21, 2022]
17 ISO(2020) ISO/TR 22758:2020 - Biotechnology - Biobanking - Implementation guide for ISO 20387. Available at: https://www.iso.org/standard/73829.html [Last accessed on December 21, 2022]
18 Hedayati M, Razavi SA, Boroomand S and Kheradmand Kia S: The impact of pre-analytical variations on biochemical analytes stability: A systematic review. J Clin Lab Anal *34(12)*: e23551, 2020. PMID: 32869910. DOI: 10.1002/jcla.23551
19 Popp D, Diekmann R, Binder L, Asif A and Nussbeck S: Liquid materials for biomedical research: a highly IT-integrated and automated biobanking solution. Journal of Laboratory Medicine *43(6)*: 347-354, 2022. DOI: 10.1515/labmed-2017-0118
20 Betsou F, Lehmann S, Ashton G, Barnes M, Benson EE, Coppola D, DeSouza Y, Eliason J, Glazer B, Guadagni F, Harding K, Horsfall DJ, Kleeberger C, Nanni U, Prasad A, Shea K, Skubitz A, Somiari S, Gunter E and International Society for Biological and Environmental Repositories (ISBER) Working Group on Biospecimen Science: Standard preanalytical coding for biospecimens: defining the sample PREanalytical code. Cancer Epidemiol Biomarkers Prev *19(4)*: 1004-1011, 2010. PMID: 20332280. DOI: 10.1158/1055-9965.EPI-09-1268
21 Betsou F, Bilbao R, Case J, Chuaqui R, Clements JA, De Souza Y, De Wilde A, Geiger J, Grizzle W, Guadagni F, Gunter E, Heil S, Kiehntopf M, Koppandi I, Lehmann S, Linsen L, Mackenzie-Dodds J, Quesada RA, Tebbakha R, Selander T, Shea K, Sobel M, Somiari S, Spyropoulos D, Stone M, Tybring G, Valyi-Nagy K and Wadhwa L: Standard PREanalytical code version 3.0. Biopreserv Biobank *16(1)*: 9-12, 2018. PMID: 29377712. DOI: 10.1089/bio.2017.0109
22 Nanni U, Betsou F, Riondino S, Rossetti L, Spila A, Valente MG, Della-Morte D, Palmirotta R, Roselli M, Ferroni P and Guadagni F: SPRECware: software tools for Standard PREanalytical Code (SPREC) labeling - effective exchange and search of stored biospecimens. Int J Biol Markers *27(3)*: e272-e279, 2012. PMID: 23032579. DOI: 10.5301/JBM.2012.9718
23 Schrohl AS, Würtz S, Kohn E, Banks RE, Nielsen HJ, Sweep FC and Brünner N: Banking of biological fluids for studies of disease-associated protein biomarkers. Mol Cell Proteomics *7(10)*: 2061-2066, 2008. PMID: 18676364. DOI: 10.1074/mcp.R800010-MCP200
24 Gandara DR, Li T, Lara PN Jr, Mack PC, Kelly K, Miyamoto S, Goodwin N, Beckett L and Redman MW: Algorithm for codevelopment of new drug-predictive biomarker combinations: accounting for inter- and intrapatient tumor heterogeneity. Clin Lung Cancer *13(5)*: 321-325, 2012. PMID: 22677432. DOI: 10.1016/j.cllc.2012.05.004
25 Nanni U, Spila A, Riondino S, Valente MG, Somma P, Iacoboni M, Alessandroni J, Papa V, Della-Morte D, Palmirotta R, Ferroni P, Roselli M and Guadagni F: RFID as a new ICT tool to monitor specimen life cycle and quality control in a biobank. Int J Biol Markers *26(2)*: 129-135, 2011. PMID: 21574153. DOI: 10.5301/JBM.2011.8323
26 Gibson CM: Moving from hope to hard work in data sharing. JAMA Cardiol *3(9)*: 795-796, 2018. PMID: 29971341. DOI: 10.1001/jamacardio.2018.0130
27 Shafto M, Conroy M, Doyle R, Glaessgen E, Kemp C, LeMoigne J and Wang L: Modeling, Simulation, Information Technology & Processing Roadmap. National Aeronautics and Space Administration, 2012. Available at: https://www.nasa.gov/sites/default/files/501321main_TA11-ID_rev4_NRC-wTASR.pdf [Last accessed on January 11, 2023]
28 Grieves M: Digital twin: Manufacturing excellence through virtual factory replication. White paper, 2014. Available at: https://www.3ds.com/fileadmin/PRODUCTS-SERVICES/DELMIA/PDF/Whitepaper/DELMIA-APRISO-Digital-Twin-Whitepaper.pdf [Last accessed on January 11, 2023]
29 Laubenbacher R, Niarakis A, Helikar T, An G, Shapiro B, Malik-Sheriff RS, Sego TJ, Knapp A, Macklin P and Glazier JA: Building digital twins of the human immune system: toward a roadmap. NPJ Digit Med *5(1)*: 64, 2022. PMID: 35595830. DOI: 10.1038/s41746-022-00610-z
30 Martinez-Velazquez R, Gamez R and El Saddik A: Cardio Twin: A digital twin of the human heart running on the edge. 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA): 1-6, 2019. DOI: 10.1109/MeMeA.2019.8802162
31 Vaskovsky A, Chvanova M and Rebezov M: Creation of digital twins of neural network technology of personalization of food products for diabetics. 2020 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR): 251-253, 2020. DOI: 10.1109/DCNAIR50402.2020.9216776
32 Voigt I, Inojosa H, Dillenseger A, Haase R, Akgün K and Ziemssen T: Digital twins for multiple sclerosis. Front Immunol *12*: 669811, 2021. PMID: 34012452. DOI: 10.3389/fimmu.2021.669811
33 An G and Cockrell C: Drug development digital twins for drug discovery, testing and repurposing: a schema for requirements and development. Front Syst Biol *2*: 928387, 2022. PMID: 35935475. DOI: 10.3389/fsysb.2022.928387
34 Sahal R, Alsamhi SH and Brown KN: Personal digital twin: a close look into the present and a step towards the future of personalised healthcare industry. Sensors (Basel) *22(15)*: 5918, 2022. PMID: 35957477. DOI: 10.3390/s22155918
35 Tellechea-Luzardo J, Winterhalter C, Widera P, Kozyra J, de Lorenzo V and Krasnogor N: Linking engineered cells to their digital twins: a version control system for strain engineering. ACS Synth Biol *9(3)*: 536-545, 2020. PMID: 32078768. DOI: 10.1021/acssynbio.9b00400
36 Stahlberg EA: Biomedical digital twins: A collaborative landscape of incredible opportunity with exciting challenges. Available at: https://www.purdue.edu/cancer-research/bigcare/Events/2022/Stahlberg_SLS_DigitalTwin-posting.pdf (Last accessed on December 21, 2022]
37 Jacobson RS, Becich MJ, Bollag RJ, Chavan G, Corrigan J, Dhir R, Feldman MD, Gaudioso C, Legowski E, Maihle NJ, Mitchell K, Murphy M, Sakthivel M, Tseytlin E and Weaver J: A federated network for translational cancer research using clinical data and biospecimens. Cancer Res *75(24)*: 5194-5201, 2015. PMID: 26670560. DOI: 10.1158/0008-5472.CAN-15-1973

Received January 13, 2023
Revised March 15, 2023
Accepted March 22, 2023