Journal for ImmunoTherapy of Cancer

Society for Immunotherapy of Cancer clinical and biomarkers data sharing resource document: Volume II—practical challenges

Alessandra Cesano,[1] Michael A Cannarile,[2] Sacha Gnjatic,[3] Bruno Gomes,[4] Justin Guinney,[5] Vaios Karanikas,[6] Mohan Karkada,[7] John M Kirkwood,[8] Beatrix Kotlan,[9] Giuseppe V Masucci,[10] Els Meeusen,[11] Anne Monette,[12] Aung Naing ,[13] Vésteinn Thorsson,[14] Nicholas Tschernia,[15] Ena Wang,[16] Daniel K Wells,[17] Timothy L Wyant,[18] Sergio Rutella[19,20]

To cite: Cesano A, Cannarile MA, Gnjatic S, *et al*. Society for Immunotherapy of Cancer clinical and biomarkers data sharing resource document: Volume II—practical challenges. *Journal for ImmunoTherapy of Cancer* 2020;**8**:e001472. doi:10.1136/jitc-2020-001472

Dr. Kotlan passed away on April 29th, 2020.

Accepted 06 October 2020

For numbered affiliations see end of article.

Correspondence to
Dr Sergio Rutella;
sergio.rutella@ntu.ac.uk

ABSTRACT

The development of strongly predictive validated biomarkers is essential for the field of immuno-oncology (IO) to advance. The highly complex, multifactorial data sets required to develop these biomarkers necessitate effective, responsible data-sharing efforts in order to maximize the scientific knowledge and utility gained from their collection. While the sharing of clinical- and safety-related trial data has already been streamlined to a large extent, the sharing of biomarker-aimed clinical trial derived data and data sets has been met with a number of hurdles that have impaired the progression of biomarkers from hypothesis to clinical use. These hurdles include technical challenges associated with the infrastructure, technology, workforce, and sustainability required for clinical biomarker data sharing. To provide guidance and assist in the navigation of these challenges, the Society for Immunotherapy of Cancer (SITC) Biomarkers Committee convened to outline the challenges that researchers currently face, both at the conceptual level (Volume I) and at the technical level (Volume II). The committee also suggests possible solutions to these problems in the form of professional standards and harmonized requirements for data sharing, assisting in continued progress toward effective, clinically relevant biomarkers in the IO setting.

INTRODUCTION: PRACTICAL CHALLENGES IN DATA SHARING FOR CLINICAL BIOMARKER DEVELOPMENT

Data sharing today enables the new science of tomorrow. It is increasingly evident that studies analyzing previously published data can achieve new discoveries, often having as much impact as the original projects, and can greatly improve medical research and benefit all stakeholders. While the companion volume to this paper, 'Society for Immunotherapy of Cancer clinical and biomarkers data-sharing resource document: Volume I—conceptual challenges', provides an overview of the surrounding framework and proposed activities of stakeholders for responsible and successful data sharing, this practical challenges volume will examine and address the more procedural details of the current hurdles to data sharing.[1] Here, we dissect those hurdles down to basic approachable elements that must be addressed to encourage better data-sharing strategies and compliance. Additionally, we put forward timely recommendations and methodological guidelines that will also remain flexible enough for adaptation to both current and future landscapes of data sharing.

For responsible data sharing to become the new norm, and to engender a lasting paradigm shift, it must first be embraced by all stakeholders and influencers. The importance of careful planning and execution of data-sharing protocols must be placed alongside any new clinical hypothesis tested. A broad range of sensible changes can be made to the current landscape of medical science in order to encourage the acceptance of data-sharing initiatives as a necessary component of scientific collaboration. In order to discuss important considerations regarding data-sharing initiatives, the Society for Immunotherapy of Cancer (SITC) Biomarkers Committee formed the Clinical and Biomarkers Data Sharing Subcommittee, which developed two manuscripts, Volumes I and II, respectively, addressing conceptual and practical challenges to data sharing.

This manuscript (Volume II) is divided into four sections, each addressing a specific group of practical challenges in data sharing. In the **Infrastructure challenges** section, we describe considerations for contracting intellectual

property (IP) and biospecimen collection associated with multicenter network infrastructure planning, considerations for the continuous validation of evolving platforms and technologies, and proposals for cloud-based computing methods that maintain required constants in communication. In the **Technological challenges** section, we describe popular data platforms that are standardizable, discoverable, searchable, and interoperable, as well as how these platforms require the development and adoption of common protocols to deliver common data elements for meaningful computation across disparate data sets. In the **Workforce challenges** section, we outline the expected knowledge and skills required of expert personnel trained to execute the design, management, technical, and operational aspects of data sharing. Finally, in the **Sustainability challenges** section, we address how the undervalued costs of data organization and sharing are borne by a small set of stakeholders and the need for these costs to be equitably redistributed across benefitting parties, in addition to how universities, research foundations, and professional societies can better foster and support data-sharing sustainability.

INFRASTRUCTURE CHALLENGES

With the ever-growing sophistication and cost of correlative science for immunotherapy clinical trials, many stakeholders have realized that the future lies in large collaborative efforts that make the best use of the breadth and depth of information generated from high-dimensional and single-cell technologies. These stakeholders include major academic centers, federal funding sources, philanthropic foundations/societies, and pharmaceutical/biotech companies. Stand-alone federal funding for laboratories is increasingly being phased out in favor of multi-principal investigator (PI) applications, with preference for those involving multiple collaborating institutions.[2 3] In addition, entities such as the Parker Institute for Cancer Immunotherapy, the Multiple Myeloma Research Foundation, the Foundation for the National Institutes of Health, and some pharmaceutical companies have engaged funding to create Networks of Excellence that bring together multiple academic centers with top expertise for collaborative streamlining of their technical and analytical pipelines. However, the infrastructure required to execute these goals needs to evolve accordingly, from virtual collaboration to the physical creation of networks with coordinated, synchronous efforts.

A recently implemented example of multicenter infrastructure for biomarker discovery is the Cancer Immune Monitoring and Analysis Centers-Cancer Immunologic Data Commons (CIMAC-CIDC) Network, funded by the National Cancer Institute's (NCI's) U24 Cancer Moonshot Initiative.[4] This grant of more than US$50 million led to the establishment of four CIMACs, which are tasked with performing a complex and comprehensive set of assays using biospecimens collected from patients undergoing immunotherapy in clinical trials under the NCI's Cancer Therapy Evaluation Program cooperative groups. The goal of this initiative is to generate rich integrated data sets that correlate treatment mechanisms (ie, multi-omic biomarkers) with clinical outcomes (ie, response to treatments, progression, and overall survival). CIDCs will maintain and centralize these integrated data sets, leading to the eventual creation of a mineable public database of well-validated, harmonized, and integrated data, with the hope that better predictive biomarkers might be discovered from these data. Establishing this pipeline for the generation, analysis, and centralized storage of this integrated data requires an unprecedented framework of infrastructure, which is organized from a series of working groups with an overseeing committee. This infrastructure enables the monitoring of all steps in the process, from trial selection and sample management, to data collection, analysis, and dissemination. This project has identified a number of obstacles met during the creation of successful data-sharing infrastructure, which are discussed below.

Contracts, IP, and ethics

Establishing contracts between institutions—including pharma/biotech companies, biobanks, and trial PIs—represents one of the principle roadblocks to data sharing. Specifically, questions arise around putting in place contracts generated only by participating centers versus the use of master agreements covering all institutions and partners in general terms, and how to distribute IP and publication deliverables from discoveries made. Different methods by which joint IP can be apportioned between contributing parties have been used, with the most straightforward method being for the involved parties to agree that any joint IP created as a result of the collaboration is jointly owned by the parties. However, this structure only works if all parties are willing to allow the unrestricted use of the joint IP by one another. Often, due to the parties' business or technical concerns, there might need to be restrictions on one or both party's use of joint IP, which could be addressed in different ways. For example, all of the IP rights could be assigned to one of the parties, which then would grant a license, limited as dictated by business or technical concerns, to the other party. Alternatively, both parties can be considered joint owners of all the joint IP, with each party agreeing to certain restrictive conditions on their use or disclosure of it. Special considerations exist when one of the parties, often an educational institution, receives government funding, since in that case the government may have rights to use IP (and possibly any underlying contributed IP) created as a result of government funding. In that case, as part of the agreement, the IP created with government funding should be available without overly burdensome restrictions, taking into account any requirements that apply to particular IP because of government funding.

If the network agrees on collaborative terms, implying shared authorship and IP rights, some rules may conflict with existing bylaws of established cooperative groups. Even if and when a common language of defined terms is finally agreed to by all parties, amendments to existing clinical protocols and consent forms need to be considered to ensure that samples can be shared in accordance with local institutional review boards (IRBs). One recommendation is to engage legal and contractual teams at each site together, rather than individually, to facilitate mutual agreements. As this includes contract offices, IP offices, material transfer agreements, IRBs, exchanges between PIs, lab testing sites, and biobanks, the networks of legal interactions between these different facilitators and stakeholders must be planned as early in the development process as possible. In addition, the language used in clinical protocols to perform correlative assays and to share resulting biospecimens for a study should allow flexibility in assays performed, and include provisions in informed consent documents for the future use of samples in later experiments or analyses, including testing of novel biomarkers.

Specimen collection

A network may depend on central biobanks or on local site processing, each having their own rules for sample collection, processing, storage, and sharing. Since an important step in biomarker validation is the evaluation of preanalytical factors that may affect assay performance (including specimen collection, handling, and processing variables), standard operating procedures for controlling specific biomarker development steps are essential and, to this effect, guidelines have been recently developed by CIMACs.[5 6] Banks and networks receiving specimens need to coordinate their laboratory information management systems and ensure end-to-end quality assurance, with 24/7 monitoring of storage conditions and temperature-controlled shipping containers. Important details to consider include codes used for the original specimens versus those used by the network, as well as the development of common vocabulary dictionaries for metadata. The latter is particularly important since metadata summarizes basic information about the data, thus making finding and working with particular instances of data easier. In addition, establishment of priority on sample use by individual institutions versus the network should be explicitly and clearly defined. The recommendations for all of these variables are summarized in an umbrella collection protocol by CIMAC, assembled by experts in assays and biobanking, and which will help investigators properly plan in advance.[7]

Some novel assays, such as single-cell high-dimensional immune profiling, tend to be dependent on high-quality materials (eg, biospecimens), typically extracted from fresh patient tissues.[8 9] Bringing surgeons, endoscopists, interventional radiologists, pathologists, and scientists together is critical to ensure timely collection of samples with minimal time to processing. Education and interest/involvement from all parties is key, and identifying such partners is vital to the success of a collaboration. However, logistical issues must also be considered, such as the need for runners (ie, biobanking technicians) and the timely delivery of specimens from operating rooms to analytical laboratories.

Equipment and conduct of research

A major reason for performing collaborative research in multiple centers is the high cost of state-of-the-art equipment. Often, these networks may choose to centralize some assays to make them more cost-effective and reduce the need to standardize protocols across different labs. However, the throughput of high-dimensional multiparametric technologies is often limiting, and most networks will need to have redundancy (eg, standards and controls) in place, to ensure that a consensus can be reached in how to interpret data generated from different experiments at different sites measuring similar metrics.

At the heart of multi-center science lies the harmonization of assays and platforms, ensuring that results can ultimately be compared. The CIMACs have spent the past 2 years benchmarking their respective assays in serial rounds of proficiency testing for multiplex immunohistochemistry (IHC), whole exome sequencing, and RNA sequencing (RNA-seq), among other assays.[5 6] A particular challenge stems from differences in platforms at all levels, from antibody clones and reagents used to machines and interpretation algorithms. Streamlining these technologies as much as possible is helpful, despite each site typically already having optimized protocols in place that are difficult to modify or replace.

Although efforts in protocol standardization are critical and should lead to vast improvements in workflows, it is also important to align expectations and to understand the limitations of harmonizing assays, especially since not everything can be tested at once. The success and details of these harmonization efforts will soon be reported, and recommendations on best practices will also be provided.[6] To ensure their relevance, assay comparisons for standardization cannot be a 'one-and-done' event, and they will, in fact, need to be continuously validated and reassessed during the process of such research.

Progress in any collaboration requires constant communication, through the organization of regular conference calls and presentations. These can also, however, quickly devolve into time sinks, especially during troubleshooting phases or real-time analysis of data. Despite expectations that these networks should be optimized for the continuous generation of data, another important priority is to continue to allow science to come first, which sometimes means that extra time is needed to ensure the generation of top-quality data, rather than adhering to a rigid schedule. An equally important recommendation is to allow assays to evolve and to improve from study to study, which is critical despite the temptation to lock them in at the same technical state for greater reproducibility. When technical improvements are introduced to keep up with

the cutting edge of research, provisions for bridging and validation work should be made to guarantee that older data also remain useful and comparable.

Aside from machines used for assays, computational and storage infrastructure for data storage, analysis, and sharing likely represents the biggest technology investment required for effective data sharing. Leasing space on cloud-based servers versus local servers needs to be carefully considered, as communications or exchange between them may be challenging. However, cloud computing may also provide new opportunities for novel research, since it can allow studies to be performed at larger scales, using data shared and integrated from multiple sources.[10] These benefits should be weighed against the possibility of risks associated with cloud computing, especially issues with data security and the need to consider the rights of research participants. In order to enable large-scale analyses and computing while maintaining patient privacy, Molnár-Gábor *et al* advocate a 'middle-ground' solution of federated or hybrid clouds, allowing large-scale cloud computing while restricting data access to approved individuals and institutions.[10] The utility of 'model to data'-type platforms also includes proper facilities for data storage, advanced data management, and analysis, and these platforms are poised to address a majority of infrastructure, ethical, technological, and sustainability issues associated with biomarker and clinical data sharing, without the need to modify current ethical and legal frameworks on the usage of clinical data.[11]

TECHNOLOGICAL CHALLENGES

The goal of data sharing is to provide transparency on how shared data sets were generated, and to reliably share comparable cross-study data, enabling increased statistical power for the identification of biomarkers that are associated with clinical benefit from the treatment being investigated. While the former enables reproducibility of data and comparison with similarly designed future trial outcomes, the limited availability of patient-derived trial samples emphasizes the need to plan for comparable data generation prospectively from a technology platform viewpoint. In this context, the current technology platforms used for analysis and robust biomarker generation represent an inherent challenge to providing comparable data sets at the single biomarker level, and even greater challenges to meeting the current and ever-growing demand for multiplex analysis of patient-derived samples for observations of the intrinsic and varied complexity of tumor-immunobiology.

With the goal of increasing the comparability of patient-derived biomarker data originating from the same or similar technology platforms, for any individual technology, in-depth knowledge on the sensitivity, specificity, stability (including positive and negative controls used), the acquisition platform, and the analysis algorithm is essential. In addition, planning on the format of data set outputs needs to be considered well in advance, early during the trial design process. Data set formats must be designed to be compatible with those emerging from similar studies to avoid the need for later conversion, which may not be financially feasible or even altogether technically possible. Currently, a number of efforts to achieve efficient data comparison are in progress, including individual as well as joint retrospective bridging studies, in addition to attempts at the development of standardized harmonization protocols for leading technology-derived assays. Below, examples of such efforts are discussed in further detail.

Variabilities of immune checkpoint expression status across different immunohistochemistry platforms

One of the most prominent and timely examples in immuno-oncology (IO) that also illustrates the critical need to harmonize platform-specific technologies is the group of antibody-based therapies called checkpoint inhibitors targeting the programmed cell death protein-1 (PD-1)/programmed death-ligand 1 (PD-L1) axis. Currently, there are six different checkpoint inhibitors approved by the US Food and Drug Administration (FDA) that have demonstrated a positive association between clinical outcome and the expression of the PD-L1 protein within the tumor microenvironment.[12 13] In patients with non-small cell lung cancer, observation of PD-L1 expression on tumor cells (using IHC technology) has been associated with clinical benefit, and has thus resulted in the development of companion/complementary diagnostic (CDx) tests designed to evaluate the PD-L1 status of each patient prior to treatment with anti-PD-1 or anti-PD-L1 checkpoint inhibitors.[14] However, multiple proprietary PD-L1 IHC assays were developed using different primary antibody clones, IHC platforms, protocols, sensitivities, and scoring methods or algorithms. Despite using the same overall technology and target (PD-L1 protein expression) the different permutations across individual assays has led to distinct staining properties and patterning for each, raising concerns about direct comparability and interchangeability of derived findings and data sets.[15] As a result, two different waves of supplemental bridging studies conducted in a precompetitive setting using both samples from clinical trials (Blueprint Phase I Project[16]) and from real-world settings (Blueprint Phase II Project[17]) have been performed. Both of these studies concluded that, with the exception of one assay, the different assays and antibody clones used showed comparable results for PD-L1 expression on tumor cells (although with some differences in numbers of positive cells detected per clone used), but showed rather poor concordance for PD-L1 expression scoring on immune cells.[16 17]

While IHC remains a highly practical and cost-effective diagnostic and prognostic method, this single-marker method cannot tell the whole story of the complex immune microenvironment. Emerging multiplex IHC and immunofluorescence technologies are promising in the field of cancer immunotherapy. Unlike conventional

IHC, which only allows the labeling of one single marker in a tissue sample, multiplex IHC is able to detect multiple markers from a single tissue sample while providing comprehensive information about the cell composition and spatial arrangement. The demand for multiplex IHC technologies that output standardized data from harmonized protocols will likely increase as it is further demonstrated that this methodology has the potential to characterize mechanisms of tumor resistance and escape in patients.[18]

Single-cell RNA-seq sample preparation and technologies to manage complex and difficult-to-harmonize data sets

In line with the increased demand for multiplexing, RNA-seq technologies have emerged as a useful tool to address complex, multi-pathway tumor-immune regulation mechanisms, and as a complement to IHC-based protein detection data sets. Beyond the already information-dense data sets generated from regular bulk RNA-seq methods, the advent of single-cell technology (scRNA-seq) offers tremendous opportunity toward greatly increasing resolution in detection of the full gamut of immune cell subtypes and simultaneous characterization of their heterogeneous functional profiles. The data sets associated with scRNA-seq technologies have the potential to be significantly larger than previous RNA-seq technologies. There are technological challenges presented by the many current platforms available for scRNA-seq, with a number of factors impacting sensitivity and operational applicability.[19] For example, the selection of the initial protocol for transcribing mRNA into cDNA (eg, full-length sequencing vs expressed sequence tag (EST) sequencing) results in differing sensitivities for less abundant transcripts.[19] In addition, the resolution of scRNA-seq results can be impacted by different cell sorting methods such as fluorescence-activated cell sorting or droplet-based microfluidics, and by the number of cells acquired, often varying from sample to sample, and between individual patients.[20–25]

For cellular subtypes of very low abundance (eg, dendritic cells in peripheral blood), a presorting enrichment step may be required to characterize subpopulations of interest. However, once data sets are generated for sharing from these different platforms in, for instance, clinical trials demonstrating the pharmacodynamics of a given treatment or for identification of response via predictive biomarkers, bridging studies comparing them will certainly be required to assess the reproducibility and feasibility of harmonization of such data sets. Recent advances in the field have demonstrated the feasibility of integrating data sets gathered from scRNA-seq with, for example, data from single-cell assays for transposase-accessible chromatin sequencing (scATAC-seq) and with data from in situ gene expression assays.[26] The expected difficulties in integrating massive numbers of data points derived from multiple scRNA-seq studies require that data sharing must be considered early in the study design process.

Tumor mutational burden assay standardization

Another important and currently emerging theme in predictive biomarkers for IO CDx development is the genomic assessment of tumor mutational burden (TMB). TMB represents both a surrogate prognostic marker and a predictive marker for the presence of tumor neo-antigens in cancer immunotherapies across multiple cancer types.[27] On June 16, 2020, the FDA granted accelerated approval to pembrolizumab for the treatment of adult and pediatric patients with unresectable or metastatic TMB-high (TMB-H) solid tumors, defined as those harboring ≥10 mutations/megabase as determined by an FDA-approved test (ie, the FoundationOne CDx test), who progressed following prior treatment and who have no alternative treatment options.[28] The direct measurement of tumor peptide antigens via methods such as mass spectrometry, and the ability to load these peptides onto major histocompatibility complex (MHC) molecules is not yet suitable for routine clinical application. However, in current clinical trials, multiple sequencing-based TMB technological platforms and panels have to date been reported to accurately quantify TMB. Francello *et al* investigated these platforms and concluded that there is a 'need for standardization of TMB quantification and reporting' in order for clinical trial TMB results to be compared for assessment and clarification of their decision-enabling potential.[29] Not surprisingly, and similarly to scRNA-seq sequencing platforms, multiple factors including library generation and sequencing, sample quality, sequencing depth, and algorithm development may generate significantly different TMB data sets across independent studies. Currently, two organizations, Friends of Cancer Research and Quality Assurance Initiative Pathology, are coordinating and proposing the standardization of TMB assessments to enable reliable and reproducible patient results.[30]

Ring studies for data standardization and lessons learned from flow cytometry technologies

Flow cytometry is an example of a technology already known for presenting challenges in data collection and analysis. This technology remains important for immuno-oncologists, as it offers the advantage of multiplexing and quantification, placing it as an indispensable tool for immune phenotyping of both the blood and tumorous patient compartments. One of the main observed challenges of flow cytometry has been the generation of comparable data across studies and across clinical trials using automated sample analysis. The individual gating criteria, as well as the signal-to-noise ratio of discrete cellular populations, may have critical impacts on the variability of data sets even in instances using the same detection methods, antibody clones, labeling fluorochromes, and sample preparation.[31] One of the data standardization pioneers offering collective ring studies for academia and industry, as well as cellular reference samples for controls, is the Association for Cancer Immunotherapy Immunoguiding Program.[32]

Another major challenge posed by this technology lies in the identification and quantification of rarer antigen-specific T cell subsets. A recent comparison of peptide-MHC multimer-binding T cells from 28 laboratories using different automated gating tools concluded that the automated gating algorithms tested scored similarly to manual central gating by detecting these cell populations in the range of 0.0005%–0.0001% of total lymphocytes.[33] None of the tested tools, however, could be fully automated, as data outputs required user-based manual decision-making. Still, a careful preselection of which tools and technologies will be used is paramount, particularly if clinical trial samples are to be analyzed using this technology. Standardization of immune phenotyping protocols, including which markers to use and how to classify immune subsets, together with standardization of gating strategies are needed for data set integration, similarly to what has been successfully done, for instance, by the EuroFlow Consortium in the field of hematologic malignancy diagnosis, prognosis, and therapeutic response prediction.[34]

Comparable clinical data elements as precursors to effective data harmonization

The sharing of comparable data sets has been discussed above as an important development that will promote the discovery of robust biomarkers. Several factors pertaining to study design impair the development of biomarkers of clinical response, including small sample sizes and observations reported too early for patients to exhibit complete or lasting responses to the agent(s) tested. These issues warrant the sharing of comparable clinical data sets, in order to assist in the establishment of well-supported prognostic biomarkers and patterns of response to IO therapies. It is also crucial that reference data sets are generated using patient cohorts receiving non-IO, standard-of-care therapies. As the field of IO still suffers from deficiencies in the consistency of expected response criteria among studies, the clinical component of shared data sets does not always report comparable clinical features, making it very difficult to correctly pool and harmonize available data sets.

Clinical response criteria and end points are a highly debated and evolving area in the field of IO. While investigators continue to favor the use of Response Evaluation Criteria in Solid Tumors (RECIST),[35] they are also considering the other models, including the consensus-based criteria for response to immunotherapy (iRECIST),[36] immune-related response criteria (irRC),[37] and immune-modified RECIST (imRECIST).[38] However, others use measures such as overall survival and 6-month progression-free survival as a surrogate of success for IO treatments.[39 40] Another challenge is that response criteria are cancer type-dependent and differ across diseases investigated (eg, pancreatic cancer vs melanoma and glioblastoma vs melanoma).[41] Two growing SITC-supported international assemblies, the TimIOs initiative and SITCure, are collectives working at forming long-lasting partnerships with both industry and academia, with the aim of assembling, pooling, and harmonizing gene expression and clinical data sets from patients receiving immunotherapies during clinical trials.[42] Their common goals are to reveal robust biomarkers of durable patient responses for the establishment of guidelines for treatment, along with treatment timelines that ensure that durable, long-lasting responses can be achieved, while minimizing possible side effects and the societal and financial burdens associated with unsubstantiated prolongation of treatments.

The sharing of common clinical data elements that can be used to establish pan-cancer immunotherapy response criteria is as critically important as any other biomarker-based feature that can be extracted from harmonized clinical trial data sets. The solving of this particular hurdle may require combined meta-analysis and artificial intelligence approaches, which will only be made possible if shared clinical data sets are complete and produced to present comparable common data elements.

In summary, the aforementioned examples underscore that efficient data set sharing requires careful consideration of the generation of high-quality raw data across technology platforms early during study or trial design, in addition to the careful planning and design of harmonizable and sharable data output deliverables. This becomes especially important when clinical pharmacodynamic biomarkers are generated to support dose selection, to provide proof of mechanism, or to identify predictive response biomarkers. In this context, the generation of comparable data sets for sharing is paramount for the eventual development and delivery of more rapid and efficient treatment opportunities for patients with cancer. There are vast numbers of studies and trials currently using multiplexing and multi-omics technologies that are generating unprecedented volumes of data, all with the singular hope of identifying simple, robust, and economical biomarkers for CDx development. These biomarkers will ideally be able to stratify patients and responses to cancer immunotherapy, but will require intermediate validation preceding standardization for routine clinical use.[43] This exemplifies the critical need for key stakeholders to subscribe to technological standards that can be applied across different trials to ascertain that both biomarker and clinical data outputs will be comparable and usable for secondary studies, bridging studies, and meta-analyses that provide the opportunity to statistically power up initial findings for robust biomarker discovery.

WORKFORCE CHALLENGES

The establishment of a better space for secondary biomarker data interrogations by making clinical trial data sets more accessible to the broader research community would require the implementation of a data standards workflow process that would allow data sharing to be undertaken in a responsible manner. At the same time, however, the implementation of such a process could present several additional complex challenges

related either to the nature of the stakeholders involved in providing such data (eg, sponsors and funders, clinical trialists, and regulatory authorities as discussed in the companion Volume I to this manuscript[1]), or to the quality and maintenance of the data committed to sharing via a central repository. Hence, in order to ensure that data sharing becomes meaningful, any prospective biomarker data repository should abide by a data standards workflow process. For this process to be successful, it should meet the following two requirements: 1) personnel encompassing a broad range of expertise to enable an end-to-end workflow in a seamless manner, and 2) data input must be provided in a manner that is usable and sharable.[44] Personnel with a broad range of expertise should provide the following:

- Regulatory oversight—Ensure that data deposited for sharing can be legally and ethically shared. Patient data should be encrypted and anonymized to eliminate any traceability to the source of origin and any identifiable data. Moreover, by maintaining continuous contact with regulatory authorities worldwide, personnel can ensure that core requirements and practices supporting the responsible sharing of clinical trial data that can be harmonized are met. The companion Volume I to this manuscript discusses the protection of patient privacy in greater detail.[1]
- Scientific oversight—Confirm the validity of the parameters measured to enable meaningful interpretation. As described above, in most IO trials, both cellular and soluble multiplex and multi-omics data are collected from the blood and tissue samples. The inputting of such voluminous data sets into a central repository designed for data sharing can be operationally, logistically, and infrastructurally challenging but crucially important to foster new discoveries.
- Bioinformatic oversight—Ensure that methods applied to data are comparable or standardized.

Considering that multi-omic data can be derived from many different sources, one must ensure their comparability among multiple data contributors. In the preceding **Technological challenges** section, examples were provided where, even when common platforms are used in, for example, flow cytometry, data quality control and analysis can result in different outputs and interpretation. This aspect becomes even more challenging when different assay platforms are used by the end user. For instance, when enumerating $CD8^+$ T cell content of a tissue, the tissue preparation conditions, the antibody clones used for staining, the tissue areas used for enumeration, and the way that tissues are scored can all lead to completely different results. Whether the scoring is undertaken in a quantitative, qualitative, digital, or manual manner can also cause discordance between different data inputs. To enable meaningful data usage, the following must be defined:

- Type of data to be deposited—By predefining the minimum amount of sample data that can or should become available for deposition, one can ensure that comparisons across several trials can be possible. Such an approach can be guided by already ongoing efforts to develop standardized biomarkers and assays, as is the case with the Partnership for Accelerating Cancer Therapies initiative that aims to provide a systematic approach to immune and oncology biomarker investigations in clinical trials.[4]
- Comparable data—As already highlighted, the data generated for deposition must abide by particular standards of execution. Whether this refers to specific technologies or to platforms, producers and users of the data must ensure that similar parameters are measured in correlative biomarker analyses. The preceding sections provide further insights as to how this can be achieved through biospecimen preparation and careful selection of technologies and data output formats that are akin to other data sets for downstream harmonization. Only then can cross-trial data comparisons result in meaningful conclusions. The availability of treatment and outcome information in integrated data sets from clinical trials, including data on patient demographics, tumor characteristics, safety, and efficacy, will ensure that benefits of data sharing are maximized.
- Format of the input data—Certain format specifications must accompany all data prepared for input. Such specifications will enable the swift integration of new usable data sets into the repository with minimal need for intervention by data curators. Adherence to standardized format specifications can also ensure that respective input data sets are readable and comprehensive and that those containing common data elements can easily be used for cross-trial comparisons. Moreover, standardized formats will minimize error introduction, and thus enable prompt availability of accurate information to secondary users. It is imperative that such standards and specifications are clearly disseminated to the data providers for application. An excellent example can be seen in the cBioPortal for Cancer Genomics, whereby a standardized bioinformatics workflow was developed, offering continuous support toward integrative biomarker analysis.[45 46]
- End-user agreements—Similar to other data-sharing platforms, the terms of use must be agreed on by all stakeholders. Considering the regulatory sensitivities imposed by such clinical data sets, there needs to be a clear mandate and guidance ensuring that all aspects of the sharing of clinical trial data are undertaken responsibly and conducted ethically. Such considerations should adopt guidance and recommendations made by regulatory bodies, either local investigators and IRBs or national/international regulatory bodies (eg, the FDA). Moreover, IP associated with the data provider must not hinder data sharing to the depository (as discussed in the companion Volume I to this manuscript).
- Access to deposited data sets from the European Union (EU) and the US—In 2018, the EU implemented the

General Data Protection Regulation (GDPR), which covers the areas of privacy, data protection, and artificial intelligence. The US legislates data privacy differently from the EU and has a variety of federal and state laws rather than one governing piece of legislation at the national level. Differences in data sharing and data distribution models between EU and US databases are exemplified by the European Genome-Phenome Archive[47] and The Cancer Genome Atlas (TCGA) data sets.[48] The former is overseen by a Data Access Committee which is responsible for reviewing applications and grants permission for access to potential users, as defined by the original informed consents. In contrast, TCGA has both an open-access data set containing information that does not pose a risk of patient re-identification as well as a controlled-access data set that contains information carrying a small risk of patient re-identification through comparing TCGA data with information from other data sets.

In summary, in order to ensure that clinical trial data sharing among the scientific community can be undertaken in a meaningful and responsible manner, a set of data standards workflow processes must be put into place. The successful implementation of such processes will require investment in bringing together appropriately trained personnel possessing a broad range of relevant expertise that ensures that all standards associated with deposition of usable data sets into a centralized repository are met. Integral to the success of such efforts is the quality of shared data with common standard data elements to ensure usability and harmonization for secondary data interrogation.

Expert personnel training for data management and sharing

Efforts aimed at increasing clinical trial data sharing will yield poor results if there are too few scientists who possess the expert knowledge required to generate or use comparable shared data for secondary analyses. An adequately sized workforce that is expertly trained in the operational and technical aspects of data sharing is thus essential. Within traditional clinical research education, the introduction of mandatory courses and course modules that specifically educate future investigators and personnel in all areas of data sharing will be valuable to ensure an adequately sized workforce. Educational programs could offer courses on the correct generation of sharable data to quantitative scientists and data scientists, with key in-class or online modules offered to medical students and clinicians slated to become or work closely with trialists. An example of these types of efforts include a freely accessible 'Research Data Management and Sharing' course already offered by the University of North Carolina at Chapel Hill and the University of Edinburgh. These types of courses are offered by a number of online providers, and can be taken for credit or audited, and thus offer training to trainees from lower-income or developing countries where financial support may be limited.[49]

In the future, international bodies that fund the training of the clinical trial workforce could make training researchers in data sharing another core component of their initiatives. Governmental funding agencies and foundations that sponsor medical research and training could also enlarge the scope of their programs to provide support for training on clinical trial design, with a focus on planning and implementation of data sharing. The Wellcome Trust, the National Institutes of Health, and the Bill and Melinda Gates Foundation are three examples of funders that impose a mandate for open-access publication and data sharing from research they support, and these and other funders could also add even stronger incentives by supporting data-sharing training of scholars.[32] Furthermore, stakeholders, including large pharmaceutical companies, could contribute state-of-the-art, hands-on training in data sharing, simultaneously increasing education on the risks associated with data sharing (ie, IP, regulatory concerns, and patient confidentiality). Other stakeholders, including professional societies such as SITC, have already held focused workshops on data sharing and have generated summarizing publications to educate leading industrial and academic researchers on current viewpoints, expectations, state-of-the-art technologies, and exemplary data-sharing platforms that may be adopted at large to shift modalities for efficient and standardized data sharing.[4] With the shared goal of creating and fostering a workforce that has the skills and knowledge to manage the operational and technical aspects of data sharing, training opportunities with clear guidelines and incentives can and should be provided by universities, funders, companies, and professional societies.

SUSTAINABILITY CHALLENGES

For data sharing to be successful, it must be sustainable, meaning it must be performed using a model where the costs of maintaining data and data-sharing resources can be equitably recuperated.[32] This challenge is exacerbated by the fact that the size of some forms of biomarker data sets are growing rapidly, and in some cases, analysis of a single sample can generate hundreds of gigabytes of data.[50] Currently, there exist a number of challenges in the implementation of sustainable data sharing, which we outline below.

Lack of understanding of data-sharing costs

To date, only a single comprehensive study has been undertaken to assess the costs of sustainable long-term data sharing.[51] While this study was important, substantially more work in this field is needed to fully understand data-sharing costs and how, precisely, overall costs are dependent on predetermined distinct variables of the data slated for sharing. Any future data-sharing landscape analysis should also address the following costs, and how these costs depend on the extent, complexity, and types of the data and data sets being shared:

Table 1 Recommendations to address practical challenges in clinical and biomarker data sharing

Challenge	Recommendation
Infrastructure	Early planning of the interactions and common technology between legal/contractual teams and other technical project architects/regulators to facilitate mutual agreements and enhance the clarity of informed consent documents
	Educating key medical/technical personnel involved in handling biospecimens to ensure timely collection and processing of samples
	Shared cloud-based storage space with real-time access and supercomputers in academic centers (with HIPAA compliance and resilience) to allow multi-core computational analyses that can be accessed by multi-center collaborators
Technology	Selection of standardizable technological platforms for generation of comparable data
	Use of supplemental bridging/ring studies to compare data-generating platforms and assess reproducibility and feasibility of data output harmonization across technologies
	Establishment of patterns of patient response profiles to guide future response criteria and trial end points
Workforce	Implementation of a data standards workflow process that allows data sharing to be meaningful and undertaken in a responsible manner
	Availability of personnel encompassing a broad range of expertise to enable an end-to-end workflow, including well rounded oversight of regulatory, scientific, curation, and bioinformatics aspects of research
	Targeted and well-supported training of expert data planning and data management personnel
Sustainability	Creation of data-sharing models where the costs of maintaining data and data-sharing resources can be better acknowledged and equitably distributed across end users
	Better defined cost factors, including required human resources for data sanitization and organization for comparability, in addition to infrastructure costs for storage and transfer
	Bioinformatics tools used to read raw data files must be available long-term, and reliable readability tools should be maintained and provided in containerized formats
	Increased recognition by academic promotion committees to incentivize data sharing
	Publishing journals encourage data sharing whenever legally and ethically possible, according to Findable, Accessible, Interoperable, and Reusable (FAIR) guiding principles [56]

HIPAA, Health Insurance Portability and Accountability Act.

- Human resource costs, notably including the costs of responding to incoming data-sharing queries and requests, and personnel required for the facilitation of data sharing.
- Data sanitization and organization costs associated with making the data usable and comparable to secondary users, which requires common data elements for data set harmonization strategies.
- Infrastructure costs, including the costs of data storage and transfer to secondary users.

Unequitable costs of data sharing

Currently, the cost categories outlined above are typically supported by stakeholders or providers generating the data, and few mechanisms exist through which secondary data users may financially support the ongoing maintenance of the important data resources on which their secondary analyses or meta-studies are based.[32 52] For lasting sustainable data-sharing models, it is imperative that end-user costs be distributed more equitably. Funding mechanisms through which researchers can support data that they wish to use should be developed. Any such mechanism would likely need to be overseen by an impartial body that would ensure fair and unbiased data access to secondary users. One benefit of such a body would be the centralization of data-use metrics to identify the features of data sets having the most utility to the biomedical research community. Importantly, shared equity of the costs of data access should not become a barrier to sharing in itself, and non-profit and/or governmental funding mechanisms to support access to data for low-resource researchers must be developed.

'Dependency hell' in bioinformatic processing tools

Unlike most forms of clinical trial data, the raw state of biomarker data is typically an unprocessed, non-human-readable format that must be subsequently analyzed and converted to generate features making downstream analyses possible. Such raw formats include .bcl, .fastq, and .bam (sequencing); .nd2, .scn, .liff, and .zvi (imaging); and .fcs (flow cytometry). Historically, bioinformatic tools used to analyze these data have typically been developed by academic research labs, and their continued maintenance and improvement is thus not guaranteed. Additionally, cascading problems may arise when shared packages

depend on multiple different and incompatible versions of the same tool, a phenomenon sometimes called 'dependency hell'. Useful, sustainable data sharing for biomarker data requires that compatible versions of tools be available and maintained.[53] One approach to address this problematic area is for the community (including academic journals) to insist that new data-processing tools be provided in a containerized format, such as Docker or Si.[54] Furthermore, long-term, centralized storage of containerized tools should be incentivized and ideally supported by a governmental funding agency. The embrace of containers for bioinformatic tools will help prevent 'dependency hell' and enable sustainable, useful data sharing of biomarkers.[55]

CONCLUSIONS

The SITC Clinical and Biomarkers Data Sharing Subcommittee analyzed the current hurdles impeding efficacy and made recommendations to set standards for data sharing in IO biomarker and clinical data sharing. The subcommittee's recommendations to address the practical challenges described in this manuscript are summarized in table 1. Priorities include early planning of legal interaction networks, cloud-based data-sharing strategies to facilitate data access and analysis by all core members, ensuring that projects are managed by personnel with expert know-how of data standards workflow practices, the application of appropriate standardizable technologies and unifying protocols that are continuously reassessed during their evolution, and that raw, usable, and comparable data outputs and storage methods be in place, with lasting and containerized bioinformatic algorithms that can continuously be used for their transformation and analysis.

The field of IO is rich in resources providing the means to treat cancer. However, there is currently no centralized database that hosts immuno-genomic data from studies involving immune checkpoint blockade, which would be a valuable resource for the IO community. There is also a need to develop next-generation computational algorithms that allow the extraction of clinically useful information from the huge amounts of data being generated using advanced molecular and cellular tools. These efforts will be critical to ultimately enable the development of precision IO treatment. The field of human genomics mapping has already been challenged by many of these questions for the last 20 or so years and represents an excellent resource for the IO field for questions of data management and sharing.

The ultimate goal of this work is to establish a culture of sharing clinical trial data in which effective incentives for data sharing exist and platforms for sharing clinical trial data are available, with appropriate data access models and with sufficient total capacity to meet demand. If more than one platform for data sharing exists, the different platforms need to be interoperable with adequate financial support for sharing clinical trial data, and with costs that are fairly allocated among stakeholders. In this ideal scenario, appropriate protections will be in place to minimize the risks of data sharing for all stakeholders and to minimize sharing disincentives. As a matter of course for best practices, shared clinical trial data need to be de-identified and modified in response to ongoing experience and feedback. The subcommittee set out to help to establish professional standards and to set expectations for responsible sharing of clinical trial data, together with requirements to be created by supporting organizations such as funders, medical journals, and professional societies (including SITC) as the best path forward, aiding the culture shift needed to implement responsible data sharing.

Author affiliations
[1]ESSA Pharma Inc, South San Francisco, California, USA
[2]Roche Pharmaceutical Research and Early Development Oncology, Roche Innovation Center Munich, Penzberg, Germany
[3]Department of Medicine, Tisch Cancer Institute, Icahn School of Medicine, New York, New York, USA
[4]Roche Pharmaceutical Research and Early Development Oncology, Roche Innovation Center, Basel, Switzerland
[5]Sage Bionetworks, Seattle, Washington, USA
[6]Roche Pharmaceutical Research and Early Development Oncology, Roche Innovation Center, Zürich, Switzerland
[7]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts, USA
[8]Department of Medicine, Division of Hematology/Oncology, University of Pittsburgh School of Medicine and Melanoma Center at UPMC Hillman Cancer Center, Pittsburgh, Pennsylvania, USA
[9]National Institute of Oncology, Budapest, Budapest, Hungary
[10]Department Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden
[11]CancerProbe Pty Ltd, Prahran, Victoria, Australia
[12]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada
[13]Department of Investigational Cancer Therapeutics, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[14]Institute for Systems Biology, Seattle, Washington, USA
[15]Department of Medicine, Division of Hematology/Oncology, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA
[16]Allogene Therapeutics, South San Francisco, California, USA
[17]Parker Institute for Cancer Immunotherapy, San Francisco, California, USA
[18]Biolojic Design Inc, Cambridge, Massachusetts, USA
[19]John van Geest Cancer Research Centre, Nottingham Trent University, Nottingham, UK
[20]Centre for Health, Ageing and Understanding Disease (CHAUD), Nottingham Trent University, Nottingham, UK

Twitter Aung Naing @AnaingMD and SITC @SITCancer

Acknowledgements The authors would like to dedicate this manuscript to the memory of Beatrix Kotlan, who served on the SITC Biomarkers Committee and made significant contributions in authoring this work. The authors acknowledge SITC staff for their contributions, including Ben Labbe, PhD for medical writing support, and Angela Kilbert, Lionel Lim, Sam Million-Weaver, PhD for project management and editorial assistance. Additionally, the authors wish to thank the society for supporting the development of the manuscript.

Contributors AC and SR, the Chairs of the SITC Clinical and Biomarkers Data Sharing Subcommittee, provided guidance on the manuscript structure and content as well as leadership of the manuscript development group. SG, BK, MK, AM, NT and EW served as section leads for the manuscript development teams. All authors

actively contributed to the manuscript development through providing content, critically reviewing drafts, and advising on additions and changes throughout the process. All authors have read and approved the final version of this manuscript.

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 AC—Employee: ESSA Pharma; Consulting fees: Refuge Bio, Arch Oncology, Qognit, Nanostring. MAC—Employee: Roche Diagnostic GMBH; Stakeholder: Roche; Patent: 10878NDR. SG—Consultancy/advisory fees: Merck, NeonTherapeutics, OncoMed; Research funding: Agenus, Bristol-Myers Squibb, Genentech, Immune Design, Janssen R&D, Pfizer, Regeneron, Takeda. BG—Employee: Hoffmann La Roche; Stockholder: Roche. VK—Employee: Hoffmann La Roche; Stockholder: Roche; Patent: EP3221355A1. JMK—Grant funding: Prometheus, Merck; Personal fees: Array Biopharma, Bristol-Myers Squibb, Novartis, Roche; Grants and personal fees: Immunocore. EM—Director/shareholder: CancerProbe Pty Ltd. AN—Consulting fees: CytomX Therapeutics, Novartis, Kymab, Genome; Contracted research: National Cancer Institute, EMD Serono, MedImmune, Healios Onc. Nutrition, Atterocor, Amplimmune, ARMO Biosciences, Eli Lilly, Karyopharm Therapeutics, Incyte, Novartis, Regeneron, Merck, Bristol-Myers Squibb, Pfizer, CytomX Therapeutics, Neon Therapeutics, Calithera Biosciences, TopAlliance Biosciences, Kymab, PsiOxus; Travel accommodation: ARMO Biosciences; Partner contracted research: Immune Deficiency Foundation. DKW—Scientific co-founder, equity holder, and paid advisor: Immunai. TLW—Employee/stockholder: Biolojic Design. JG, MK, BK, GVM, AM, SR, VT, NT, and EW—Nothing to disclose. SITC Staff: AK, BL, LL, and SMW—Nothing to disclose.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Aung Naing http://orcid.org/0000-0002-4803-8513

REFERENCES

1 Rutella S, Cannarile M, Gnjatic S, *et al*. Society for immunotherapy of cancer clinical and biomarkers data sharing resource document: volume I—conceptual challenges. *J Immunother Cancer*2020;8;e001389.
2 Paolino AR, Lauf SL, Pieper LE, *et al*. Accelerating regulatory progress in multi-institutional research. *eGEMs* 2014;2:10.
3 Steiner JF, Paolino AR, Thompson EE, *et al*. Sustaining research networks: the twenty-year experience of the HMO research network. *EGEMS* 2014;2:1.
4 Butterfield LH, Disis ML, Fox BA, *et al*. SITC 2018 workshop report: Immuno-Oncology biomarkers: state of the art. *J Immunother Cancer* 2018;6:138.
5 Cancer Immune Monitoring and Analysis Centers & Cancer Immunologic Data Commons. Assay SOPs. Available: https://cimac-network.org/assay-sops/
6 Cancer Immune Monitoring and Analysis Centers & Cancer Immunologic Data Commons. Guidelines for data access/transfer and publications for correlative studies involving collaboration between the CIMAC-CIDC network and the clinical investigators/Clinical trial networks on NCI-supported clinical trials 2019.
7 CIMAC-CIDC Immuno-Oncology Biomarkers Network. Documents: CIMAC specimen collection umbrella. Available: https://cimac-network.org/documents/
8 Gubin MM, Esaulova E, Ward JP, *et al*. High-Dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful Immune-Checkpoint cancer therapy. *Cell* 2018;175:1014–30.
9 Chuah S, Chew V. High-Dimensional immune-profiling in cancer: implications for immunotherapy. *J Immunother Cancer* 2020;8:e000363.
10 Molnár-Gábor F, Lueck R, Yakneen S, *et al*. Computing patient data in the cloud: practical and legal considerations for genetics and genomics research in Europe and internationally. *Genome Med* 2017;9:58.
11 Ellrott K, Buchanan A, Creason A, *et al*. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol* 2019;20:195.
12 Gong J, Chehrazi-Raffle A, Reddi S, *et al*. Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. *j. immunotherapy cancer* 2018;6:8.
13 Shen X, Zhao B. Efficacy of PD-1 or PD-L1 inhibitors and PD-L1 expression status in cancer: meta-analysis. *BMJ* 2018;362:k3529.
14 Büttner R, Gosney JR, Skov BG, *et al*. Programmed Death-Ligand 1 immunohistochemistry testing: a review of analytical assays and clinical implementation in non-small-cell lung cancer. *J Clin Oncol* 2017;35:3867–76.
15 Ratcliffe MJ, Sharpe A, Midha A, *et al*. Agreement between programmed cell death ligand-1 diagnostic assays across multiple protein expression cutoffs in non-small cell lung cancer. *Clin Cancer Res* 2017;23:3585–91.
16 Hirsch FR, McElhinny A, Stanforth D, *et al*. Pd-L1 immunohistochemistry assays for lung cancer: results from phase 1 of the blueprint PD-L1 IHC assay comparison project. *J Thorac Oncol* 2017;12:208–22.
17 Tsao MS, Kerr KM, Kockx M, *et al*. Pd-L1 immunohistochemistry comparability study in real-life clinical samples: results of blueprint phase 2 project. *J Thorac Oncol* 2018;13:1302–11.
18 Halse H, Colebatch AJ, Petrone P, *et al*. Multiplex immunohistochemistry accurately defines the immune context of metastatic melanoma. *Sci Rep* 2018;8:11158.
19 See P, Lum J, Chen J, *et al*. A single-cell sequencing guide for Immunologists. *Front Immunol* 2018;9:2425.
20 Andreyev DS, Zybailov BL. Integration of flow cytometry and single cell sequencing. *Trends Biotechnol* 2020;38:133–6.
21 Baron CS, Barve A, Muraro MJ, *et al*. Cell type purification by single-cell Transcriptome-Trained sorting. *Cell* 2019;179:527–42.
22 Zilionis R, Nainys J, Veres A, *et al*. Single-Cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* 2017;12:44–73.
23 Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;161:1187–201.
24 Macosko EZ, Basu A, Satija R, *et al*. Highly parallel genome-wide expression profiling of individual cells using Nanoliter droplets. *Cell* 2015;161:1202–14.
25 Zheng GXY, Terry JM, Belgrader P, *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;8:14049.
26 Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;177:1888–902.
27 Samstein RM, Lee C-H, Shoushtari AN, *et al*. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* 2019;51:202–6.
28 Food and Drug Administration. KEYTRUDA prescribing information. Available: https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&ApplNo=125514
29 Fancello L, Gandini S, Pelicci PG, *et al*. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *J Immunother Cancer* 2019;7:183.
30 Stenzinger A, Allen JD, Maas J, *et al*. Tumor mutational burden standardization initiatives: recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. *Genes Chromosomes Cancer* 2019;58:578–88.
31 Gouttefangeas C, Chan C, Attig S, *et al*. Data analysis as a source of variability of the HLA-peptide multimer assay: from manual gating to automated recognition of cell clusters. *Cancer Immunol Immunother* 2015;64:585–98.
32 Lo B. Sharing clinical trial data: maximizing benefits, minimizing risk. *JAMA* 2015;313:793–4.
33 Pedersen NW, Chandran PA, Qian Y, *et al*. Automated analysis of flow cytometry data to reduce Inter-Lab variation in the detection of major histocompatibility complex Multimer-Binding T cells. *Front Immunol* 2017;8:858.
34 van Dongen JJM, O'Gorman MRG, Orfao A. EuroFlow and its activities: introduction to the special EuroFlow issue of the Journal of immunological methods. *J Immunol Methods* 2019;475:112704.
35 RECIST Working Group. RECIST. Available: https://recist.eortc.org/
36 RECIST Working Group. iRECIST. Available: https://recist.eortc.org/irecist/
37 Wolchok JD, Hoos A, O'Day S, *et al*. Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. *Clin Cancer Res* 2009;15:7412–20.
38 Hodi FS, Ballinger M, Lyons B, *et al*. Immune-Modified response evaluation criteria in solid tumors (imRECIST): refining guidelines to

assess the clinical benefit of cancer immunotherapy. *J Clin Oncol* 2018;36:850–8.
39 Wang Z-X, Wu H-X, Xie L, *et al*. Correlation of milestone restricted mean survival time ratio with overall survival hazard ratio in randomized clinical trials of immune checkpoint inhibitors: a systematic review and meta-analysis. *JAMA Netw Open* 2019;2:e193433.
40 Cottrell TR, Thompson ED, Forde PM, *et al*. Pathologic features of response to neoadjuvant anti-PD-1 in resected non-small-cell lung carcinoma: a proposal for quantitative immune-related pathologic response criteria (irPRC). *Ann Oncol* 2018;29:1853–60.
41 Nishino M, Giobbie-Hurder A, Gargano M, *et al*. Developing a common language for tumor response to immunotherapy: immune-related response criteria using unidimensional measurements. *Clin Cancer Res* 2013;19:3936–43.
42 Guerriero J, Thaxton J, Barkowiak T, *et al*. P719 a SITC-sponsored randomized clinical trial to determine criteria to guide clinicians on when to stop immunotherapy through a community-driven data Repository. *leveraging the SITC community* 2019.
43 Cesano A, Marincola FM, Thurin M. Status of Immune Oncology: Challenges and Opportunities. In: Thurin M, Cesano A, Marincola FM, eds. *Biomarkers for immunotherapy of cancer: methods and protocols*. New York, NY: Springer New York, 2020: 3–21.
44 Karim MR, Michel A, Zappa A, *et al*. Improving data workflow systems with cloud services and use of open data for bioinformatics research. *Brief Bioinform* 2018;19:1035–50.
45 Cerami E, Gao J, Dogrusoz U, *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data: Figure 1. *Cancer Discov* 2012;2:401–4.
46 Gao J, Aksoy BA, Dogrusoz U, *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;6:pl1–pl.
47 European Genome-phenome Archive. EGA: about data access. Available: https://www.ebi.ac.uk/ega/about/access
48 National Cancer Institute. The cancer genome atlas program. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
49 Liyanagunawardena TR, Williams SA. Massive open online courses on health and medicine: review. *J Med Internet Res* 2014;16:e191.
50 Prosperi M, Min JS, Bian J, *et al*. Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak* 2018;18:139.
51 Wilhelm EE, Oster E, Shoulson I. Approaches and costs for sharing clinical research data. *JAMA* 2014;311:1201–2.
52 Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res* 2015;43:W566–70.
53 List M, Ebert P, Albrecht F. Ten simple rules for developing Usable software in computational biology. *PLoS Comput Biol* 2017;13:e1005265.
54 Boettiger C. An introduction to docker for reproducible research. *SIGOPS Oper. Syst. Rev.* 2015;49:71–9.
55 Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal* 2014;2014.
56 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.