MDPI

Opinion

Open Data to Support CANCER Science—A Bioinformatics Perspective on Glioma Research

Fleur Jeanquartier [1], Claire Jean-Quartier [1,2,*], Sarah Stryeck [2] and Andreas Holzinger [1,2,3]

1   Human-Centered AI Lab (Holzinger Group), Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, 8036 Graz, Austria; f.jeanquartier@hci-kdd.org (F.J.); andreas.holzinger@human-centered.ai (A.H.)
2   Institute for Data Science and Interactive Systems, Graz University of Technology, 8010 Graz, Austria; sarah.stryeck@tugraz.at
3   xAI Lab, Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB T6G 2E8, Canada
*   Correspondence: c.jeanquartier@hci-kdd.org

Simple Summary: In this opinion paper, we advocate the reuse and sharing of data supporting cancer science. We highlight artificial intelligence methods based on open data for insight generation and present use cases in glioma research.

Abstract: Supporting data sharing is paramount to making progress in cancer research. This includes the search for more precise targeted therapies and the search for novel biomarkers, through cluster and classification analysis, and extends to learning details in signal transduction pathways or intra- and intercellular interactions in cancer, through network analysis and network simulation. Our work aims to support and promote the use of publicly available resources in cancer research and demonstrates artificial intelligence (AI) methods to find answers to detailed questions. For example, how targeted therapies can be developed based on precision medicine or how to investigate cell-level phenomena with the help of bioinformatical methods. In our paper, we illustrate the current state of the art with examples from glioma research, in particular, how open data can be used for cancer research in general, and point out several resources and tools that are readily available. Presently, cancer researchers are often not aware of these important resources.

Keywords: open science; publicly available data; open; glioma; cancer; AI; explainability

1. Introduction

What are the currently known biomarkers and cancer driver genes for a selected sub-disease? Which genetic aberrations can be used diagnostically? Have survival-associated patterns been already identified? Which overall survival can be predicted? Are there any gender and age specifics about certain cancer subtypes? Are there any targeted drug recommendations for certain genomic variations? Numerous questions are being raised regarding cancer research every day and, partly, data already exist that help find answers. Yet, it is still not generally acknowledged to reuse and share data in cancer research [1,2].

Biomolecular data types range from genomic, proteomic, and metabolomic, up to radiomic and clinical data. These include cancer-related whole genome and large-scale genomic sequencing data, copy number alterations, DNA methylation, different types of mutations, microarray data, microRNAs, RNA sequencing data, protein-protein interaction (PPI) probing, protein mass spectrometry, drug-target relationships, and further biological and pharmacological data, as well as cancer incidence, mortality rates, prevalence, and survival rates [3].

2. Open Data for Cancer Research

This section is structured as follows:

1. Why open data research?—from open to FAIR (findable, accessible, interoperable, and reusable).
2. General biomedical data providers.
3. Cancer specific data initiatives and resources.
4. Metadata for AI in cancer research.
5. Explainability and causability.
6. Fostering exchange and use cases for glioma research.

2.1. Ad 1. Why Open Data Research?

In the year of 1957, the International Council of Scientific Unions (ICSU) prepared the International Geophysical Year, amongst other reasons, to overcome the many data locks of the cold war times [4]. Recently, the ICSU was merged with the International Social Science Council (ISSC) to form the International Science Council (ISC) [5]. In the last quarter of the past century, the idea of worldwide data exchange grew, which resulted in the necessity of the standardization of metadata for exchange [6]. In the 1970s, National Aeronautics and Space Administration (NASA) had to cooperatively work with international partners to operate ground control stations, leading to the implementation of a standardized way of data exchange [7]. By now, NASA has its very own open data portal [7]. In 1995, the national academy of sciences published a report "On The Full And Open Exchange Of Scientific Data". Within this report, the committee on the geophysical and environmental data of national research council, Washington, D.C., demanded the disclosure of data and promoted open exchange between different countries [8]. It was the end of 2005 when common endeavours to collect and share the genomic analysis of 33 different cancer types with The Cancer Genome Atlas (TCGA) was launched, 2006, followed by Therapeutically Applicable Research to Generate Effective Treatments (TARGET), concerned with childhood cancer research [9] and International Cancer Genome Consortium's (ICGC) in 2008 [10]. Local initiatives followed, such as the German cancer consortium (DKTK) in 2012 [11]. Additionally, other global initiatives followed. In 2014, the Global Alliance for Genomics and Health (GA4GH) was founded, in order to enable responsible genomic data sharing. Soon after, global corporations throughout the world joined, supporting data sharing initiatives in cancer research [12].

Biomedical databases provide both open, as well as controlled, access data, depending on data type, such as for the ICGC data portal [10]. Open (access) data are data that can be used by anyone, without technical or legal restrictions. The use encompasses both access and reuse. Still, open data is less common than open access publications, which are two of many important research stages in open(ing) science [13]. AI development requires diverse, publicly available, and annotated data, in regard to quality, validation, and reproducibility. This aspect becomes more and more important, with an increased amount of data being produced every day. The recent year has proven that open science can save lives [14]. Besides open data, the FAIR principles developed as a concept to ensure the reproducibility and quality of research. FAIR does not only apply to data but also to tools and services (e.g., repositories). FAIR data makes data findable (e.g., through a digital object identifier), accessible (e.g., through repositories), interoperable (e.g., through the use of open formats and technologies), and re-usable (e.g., through adequate documentation with metadata), while still protecting individuals privacy, which is essential, in case of sensitive patient data. In order to adhere with FAIR principles, it is crucial to have access to technological solutions (e.g., repositories) but also to have discipline-specific know-how for the adequate documentation or use of metadata standards [15–17].

2.2. General Biomedical Data Providers

In the area of biomedicine, vast amounts of data are produced; meanwhile, international institutions exist that provide data and tools openly by, and to, the scientific community. There are two big institutions that provide open data for bioinformatic research, including cancer data. Famous worldwide is the National Center for Biotechnology

Information (NCBI), located in the United States. Another key player in providing resources for bioinformatic research is the European Bioinformatics Institute (EMBL–EBI), located in the United Kingdom. NCBI provides many national resources but also participates in international projects, including EMBL–EBI projects and vice-versa [18]. Additionally, EMBL–EBI provides many internationally curated, high-quality data resources, including data from teams worldwide, following a coherent strategy [7].

There are many resources available online, from old and outdated ones to highly curated, disease-specific data repositories, providing data with full open access, semi-free access, and some that require data requests to grant access. One of the most famous open access data providers is Pubmed Central (PMC), which provides many full abstracts and both free full-text publications, as well as information that links to publishers with restrictions. PubChem is a freely accessible chemical information database with information about chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations, and more. Gene Expression Omnibus (GEO) is a functional genomics data repository with querying tools and download options for array- and sequence-based data. PMC, PubChem, and GEO, among others, are services from the before-mentioned NCBI [18]. Ensembl, UniProt, Protein Data Bank in Europe (PDBe) (but also the larger content provider Europe PMC), ChEMBL, ArrayExpress (currently being migrated to BioStudies), and the Expression Atlas are some of the more famous data resources provided by EMBL–EBI [19,20]. Ensembl currently supports data from more than 50,000 genomes across the different websites. Uniprot is a comprehensive, high-quality database of protein sequences and functional information. PDBe is the European descendant of the worldwide Protein Data Bank (PDB) [21], collecting, organising, and disseminating data on biological molecular structures. ChEMBL combines chemical, genomic, and bioactivity data of drug-like molecules. ArrayExpress collects data from high-throughput functional genomics experiments. Expression Atlas makes use of ArrayExpress data. There are also joint repositories next to the worldwide PDB, such as the Consensus Coding Sequence Database (CCDS) [22] or GLOBOCAN cancer statistics, provided by International Agency for Research on Cancer (IARC), a specialized cancer agency of the World Health Organization (WHO) [23]. Smaller local, and more specific resources, are also available, such as the Chinese Glioma Genome Atlas (CGGA) [24]. PDB provides access to structural data for biologial molecules. CCDS collects high-quality annotated protein coding regions in human and mouse genomes. GLOBOCAN provides global cancer statistics for cancer control and research. CGGA is a resource with functional genomic data from Chinese gliomas. Most of these resources provide information on which data is available but also how to contribute to the projects. For instance, regarding BioStudies, which data and how to submit is described in https://www.ebi.ac.uk/biostudies/submit (accessed on 12 December 2021). There are also several imaging data repositories from EMBL–EBI, providing images of different molecular scales, ranging from macro-molecular subcellular structures, up to large tissue masses: EMPIAR, Cell-IDR, Tissue-IDR, BioImage Archive, and many more [25]. In the area of life science, one can find comprehensive lists for research data management practice, f.i. in https://github.com/elixir-europe/rdmkit (accessed on 12 December 2021). A table of data resources, with causal information in biological databases, can be found in [26]. However, disease-specific, in particular, on a certain cancer types, data availability varies. The next subsection describes cancer-specific resources. We try to summarize most important resources in Table 1 and relate to specific use case examples in Table 2.

Table 1. Overview of general biomedical and cancer-specific resources.

Name	Type of Data	Provider
general biomedical data resources		
Pubmed Central (PMC)	publications, references	NCBI (NIH)
PubChem	molecule information	NCBI (NIH)
Gene Expression Omnibus (GEO)	gene expression	NCBI (NIH)
Europe PMC	publications, references	EMBL–EBI
Ensembl	genomic information	EMBL–EBI
UniProt	protein information	EMBL–EBI
Protein Data Bank in Europe (PDBe)	protein structures	EMBL–EBI
ChEMBL	molecule information	EMBL–EBI
ArrayExpress/Biostudies	from functional genomics to a variety of study data	EMBL–EBI
Expression Atlas	gene expression	EMBL–EBI
BioImage Archive (BIA)	images of all scales	EMBL–EBI
Protein Data Bank (PDB)	protein structures	joint, worldwide
Consensus Coding Sequence Database (CCDS)	genome sequences	joint, worldwide
cancer-specific resources		
GLOBOCAN	cancer statistics	IARC (WHO)
The Cancer Genome Atlas (TCGA)	cancer genomics	NCI (NIH)
The Cancer Imaging Archive (TCIA)	cancer images	NCI (NIH)
Surveillance, Epidemiology, and End Results (SEER)	cancer incidences	NCI (NIH)
International Cancer Genome Consortium (ICGC)	cancer genomics	joint, worldwide
Catalogue Of Somatic Mutations In Cancer (COSMIC)	cancer somatic mutations	WSI (England)
cBio Cancer Genomics Portal (cBioPortal)	cancer genomics	joint, worldwide
Chinese Glioma Genome Atlas (CGGA)	glioma genomics	Beijing Neuro-surgical Institute
Pediatric Cancer Genome Project (PCGP)	cancer genomics	joint, worldwide
Cancer Cell Map Initiative (CCMI)	cancer cell maps	joint (UCSC a.o.)
kipoi	cancer genomic models	joint, worldwide

Table 2. Use case examples using *mixed* and *open* resources for cancer research with focus on glioma.

	Example USE Cases	Mixed	Open
modeling and simulation	tumor growth, migration, angiogenesis	[27,28]	[29,30]
clustering and network analysis	biomarker discovery, grading, subtype, drug and pathway analysis	[31–33]	[34–40]
radiomic analysis	diagnosis and survival	[28,41]	[42–44]
information retrieval and science communication	epidemiology, education, investigation, bibliometrics	[45–47]	[23,48,49]

2.3. Cancer Specific Data Initiatives and Resources

Regarding the topic of cancer research, there are also some disease-specific resources provided by the US National Cancer Institute (NCI). To name some of the most important ones, TCGA is available via the Genomic Data Commons Portal at https://portal.gdc.cancer.gov/ (accessed on 12 December 2021). The Cancer Imaging Archive (TCIA), also sponsored by NCI, provides radiomics data [50] via https://www.cancerimagingarchive.net/ (accessed on 12 December 2021). Radiomics data can be submitted to TCIA, following the guide in https://www.cancerimagingarchive.net/primary-data/ (accessed on 12 December 2021). The Pan Cancer Analysis of Whole Genomes (PCAWG) is one of the ICGC initiatives that provides common patterns of mutation among different cancer types. PCAWG data is available via several databases, such as the ICGC data portal but also the Expression Atlas and the University of California Santa Cruz (UCSC)'s Xena Functional Genomics Explorer [51]. For instance, differential network analysis can be applied using the Expression Atlas and PCAWG data [34]. The cBio Cancer Genomics Portal (cBioPortal) is another collaborative effort that provides open genomic data, including TCGA pancancer studies, as well as open source software for local instances [52]. Data from cBioPortal, and its pediatric-specific instance, pedcBioPortal, can be used for clustering and classification analysis [35]. The multi-institutional systems biology center Cancer Cell Map Initiative (CCMI) supports NDEx, providing data commons for biological networks [53]. To overcome the lack of data from young patients, the Pediatric Cancer Genome Project (PCGP) provides data via https://pecan.stjude.cloud/pcgp-explore (accessed on 12 December 2021) [36]. The Catalogue of Somatic Mutations in Cancer (COSMIC) can be accessed via https://cancer.sanger.ac.uk/cosmic (accessed on 12 December 2021). COSMIC is provided

by Wellcome Sanger Institute (WSI), located in the United Kingdom. COSMIC uses data from ICGC, TCGA, and others. Several other resources can be found and are discussed elsewhere [54]. Glioma-specific web resources, partly making use of data provided by these initiatives, are further described in Section 2.6. To support the scientific community, a notable example for in silico resources is Kipoi, a repository of reusable predictive genomic models, where researchers are able to both contribute, as well as reuse and compare [55]. Additionally, datasets dedicated to finding suitable AI methods are growing [42]. Generally, data sharing is named as one key limitation in AI research [56].

2.4. Metadata for AI in Cancer Research

Reports on machine learning applications in medical science often lack accessibility or reproducibility and describe only selected aspects of the models; still, trust in biomedical applications is of particular importance in medical science [57]. The clinical utility of AI applications would require the evaluation of external cohorts and documentation in online repositories [58]. Next to the challenges of finding sufficiently large, diverse, and well-annotated datasets for AI training, there is the issue of data privacy and ownership that significantly hampers model development in medicine. This aspect makes transfer learning and, moreover, federated learning approaches, based on distributed model-training to data-owners, more and more prominent [59,60]. Additionally, the EU recently published a regulatory framework on AI, to propose a list of high-risk applications, set requirements, and define specific obligations for AI users and providers of high-risk applications [61].

2.5. Explainability and Causability

Although explainable AI (xAI) has only recently become popular as a field, the problem of explainability is practically the oldest field of science and is well anchored in the philosophy of science [62]. Actually, the problem has arisen due to the great successes of statistical machine learning and the non-linear models, such as complex neural networks (deep learning), that make it practically impossible to track all steps to a result. However, this traceability is now necessary for legal reasons, and xAI is now developing a series of post-hoc models that make it possible for results of so-called black-box models to be understood, comprehended, and interpreted by the end users [63,64]. These methods can be very useful in biology, medicine, and the life sciences, e.g., [65,66]. However, in certain domains, especially in the medical domain, there is a need for causability, referring to a human model, instead of the technical approach of explainability [65,67]. Causability, introduced in reference to usability, corresponds to the measurable extent to which an explanation, resulting from an xAI method to a human expert, reaches a certain level of causal understanding, measured with the system causability scale [68], causal, in Judea Pearl's sense, as the relationship between cause and effect [69]. Understanding can be reached if explainability is mapped with causability, which requires new human-AI interfaces that allow domain experts to interactively ask questions and counterfactuals to gain insights into the underlying explanatory factors of an outcome [70], likewise supporting reproducibility [71].

### 2.6. Fostering Exchange and Use Cases for Glioma Research

Modeling brain tumor-related studies exist that allow for the **simulation** of tumor growth [29] and resection [30], making use of open data, as well as providing open source implementations to reproduce and further refine model parameters. Moreover, using open data for cancer research can support biomarker prediction [72]. With the help of the pan cancer analysis of TCGA data, the evaluation of the mRNA level of traditionally used reference genes revealed novel ones for specific cancer types [38]. Brain tumor subtype classification has been based on TCGA brain cancer multi-omics data [37].

Network analysis and clustering benefit from several open cancer resources [34,35]. The combination of various data sets and types can further lead to novel findings of signal transduction events, leading to new therapy possibilities. This has been done, for exam-

ple, in the case of using publicly available gene expression data by GEO, transcriptomic sequencing data by CGGA, and RNA-sequencing data by TCGA [40]. Another notable example for targeting cancer studies is the immune landscape of cancer [49]. An exemplary glioma-specific web resource uses raw and annotated data from several sources including TCGA, GEO, COSMIC, ClinVar, FDA, etc., for **network** visualizations [33]. Another example is described by a web resource on metabolomic data [73,74]. Metabolic data has likewise been used for molecular classification and biomarker discovery in glioma research [75]. Additionally, several metabolic alterations have been highlighted in glioma patients [32]. Another example use case is described by combining metabolic profiles with transcriptomic and proteomic data [39].

Radiomics constitutes the discipline on medical image analysis, in regard to harnessing radiomic features, which are extracted quantitative metrics, using methods, such as feature calculation, selection, dimensionality reduction, and data processing [76]. Non-invasive imaging is readily used for monitoring tumor mass and treatment resistance and can be included in patient-specific models on tumor growth and response to chemoradiation [28]. Open access tools and medical image repositories already exist to support radiomic approaches [77]. Moreover, open data is used for solving brain tumor segmentation challenges [41,42]. Classification can be based on various data types, also using radiomics [44]. The combinatory use of medical images and genomic features, described by radiogenomics, can be used for clinical outcome prediction and guiding therapy [78].

In both public and **scientific communication**, it is the goal to foster understanding [79], such as the dissemination of cancer inequities [23] or facing challenges as uncertainties [80]. Benefits of mapping and visualization are used to tackle varying informational needs [48,81]. Specific glioma gene expressions can be visually analyzed with the tool Glioblastoma Bio Discovery Portal (GBM-BioDP) [48], next to other more general cancer TCGA visualization tools [72]. Prognostic markers, as well as genetic risk factors, can be reviewed, with the help of molecular epidemiology [47]. The surveillance, epidemiology, and end results (SEER) data can be used to study risks that may occur after radiation therapy of pediatric LGG [82]. Bibliometrics can show trends for specific research topics. Figure 1 shows the growing number of published documents on open data, related to cancer, as well as open access share on publications. Bibliometric analyses related to glioma exist, which make use of Scopus, ranking both open access and closed access publications [45,46]. Data from the past years were used to report estimations on new cases and deaths globally for the upcoming year. Challenges arise for cancer registries to exchange incidence data, regarding national regulations concerned with data privacy [23]. Examples, such as the proportional increase in open access publications on glioma, illustrated in Figure 1, show a tiny, but recognizable, trend towards opening science.
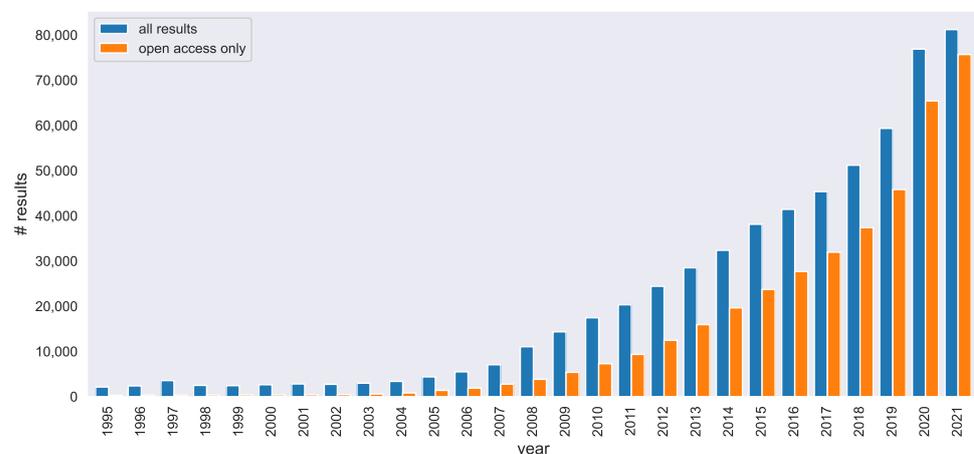
Figure 1. Pubmed Central (PMC) publication results on search for "open data cancer" per year, comparing the number of all results and filtered by open access.

3. Conclusions

Regarding all the opportunities that come with open data, as a part of open science, it is still essential to further publish data with free access.

Among top limitations are data privacy laws, technology, and lack of expertise [83]. Challenges, regarding data privacy, include re-identification risks [10]. In contrast to legal issues with privacy, computer science methods are concerned with data protection, which brings us to technological challenges and, up to now, certain limitations. The more, the better is not always true. Regarding the application of ML models in cancer, large datasets may result in overfitting and/or bias; therefore, training data sets should be diverse, as well as representative [58]. Thus, next to quantity, data quality is of particular importance, since data sets used for AI approaches require thorough curation and processing [84]. This aspect includes several factors, such as expert labeling [85], completeness, harmonization, and standardization [86], just as validation [87].

There is a discipline-specific tendency to share data openly, as is common for biology researchers but to a lesser extent for medical or pharmaceutical scientists, based on several drivers and inhibitors for sharing and using open research data, including aspects such as the researchers background and experience, intrinsic motivation, trust facilitating conditions, social influence and affiliation, expected performance, effort, requirements and formal obligations, legislation, and regulation, next to data characteristics [2].

The growth of image repositories is suggested to have a great impact on AI, with clinical relevance, in the future [56]. Unfortunately, many examples in radiomics lack openness, both in data and source code and, therefore, reproducibility. While radiomics is becoming more interdisciplinary, not only including medicine but also computer science, reports also emerge that already include accessible links to source code within the publication, such as in radiomic studies [41–43], as well as other related cancer research [29,35,52]. Another issues concern long-term financing. Examples, such as GliomaDB [33], show that small projects, with limited funding, can only offer temporary solutions. To pursue such solutions, it is essential to broaden thought beyond distribution and maintenance.

More openness across institutes will help us to exchange research with others and foster novel outreach and engagement activities. Therefore, we propose to share and reuse research output towards decoding diseases, such as cancer, together [8,9,14,25,50,72,88].

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CCDS	Consensus Coding Sequence Database
CCMI	Cancer Cell Map Initiative
CGGA	Chinese Glioma Genome Atlas
COSMIC	Catalogue of Somatic Mutations in Cancer
DKTK	German Cancer Consortium: Deutsches Konsortium für Translationale Krebsforschung
EMBL–EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
FAIR	Findable, Accessible, Interoperable, and Reusable
GA4GH	Global Alliance for Genomics and Health
GDC	Genomic Data Commons
IARC	International Agency for Research on Cancer
ICGC	International Cancer Genome Consortium
ICSU	International Council of Scientific Unioins
ISC	International Science Counsil
ISSC	International Social Science Counsil
NASA	National Aeronautics and Space Administration
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
PPI	Protein Protein Interaction
PCAWG	Pancancer Analysis of Whole Genomes
PCGP	Pediatric Cancer Genome Project
PDB(e)	Protein Data Bank (in Europe)
PMC	PubMed Central
PPI	protein-protein interaction
SEER	Surveillance, Epidemiology, and End Results
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Project
TCIA	The Cancer Imaging Archive
UCSC	University of California Santa Cruz
WHO	World Health Organization
WSI	Wellcome Sanger Institute
xAI	Explainable Artificial Intelligence

References

1. Jean-Quartier, C.; Jeanquartier, F.; Jurisica, I.; Holzinger, A. In silico cancer research towards 3R. *BMC Cancer* **2018**, *18*, 408.
2. Zuiderwijk, A.; Shinde, R.; Jeng, W. What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. *PLoS ONE* **2020**, *15*, e0239283. [CrossRef]
3. Vamathevan, J.; Apweiler, R.; Birney, E. Biomolecular data resources: Bioinformatics infrastructure for biomedical data science. *Annu. Rev. Biomed. Data Sci.* **2019**, *2*, 199–222. [CrossRef]
4. Aronova, E.; Baker, K.S.; Oreskes, N. Big science and big data in biology: From the international geophysical year through the international biological program to the long term ecological research (LTER) Network, 1957—-Present. *Hist. Stud. Nat. Sci.* **2010**, *40*, 183–224. [CrossRef]
5. Esteban, M.J.; Puppo, G. The New International Science Council–A Global Voice for Science. *EMS Newsl.* **2018**, *109*, 49. [CrossRef]
6. Goldstein, B.; Kemmerer, S.; Parks, C. *A Brief History of Early Product Data Exchange Standards*; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 1998.
7. Nicol, A.; Caruso, J.; Archambault, É. Open data access policies and strategies in the European research area and beyond. *Info@ Sci.* **2013**, *1*, 495–6505.
8. National Research Council. *On the Full and Open Exchange of Scientific Data*; The National Academies: Washington, DC, USA, 1995. [CrossRef]
9. Hinkson, I.V.; Davidsen, T.M.; Klemm, J.D.; Chandramouliswaran, I.; Kerlavage, A.R.; Kibbe, W.A. A comprehensive infrastructure for big data in cancer research: Accelerating cancer research and precision medicine. *Front. Cell Dev. Biol.* **2017**, *5*, 83. [CrossRef]

10. Milius, D.; Dove, E.S.; Chalmers, D.; Dyke, S.O.; Kato, K.; Nicolas, P.; Ouellette, B.F.; Ozenberger, B.; Rodriguez, L.L.; Zeps, N.; et al. The International Cancer Genome Consortium's evolving data-protection policies. *Nat. Biotechnol.* **2014**, *32*, 519–523. [CrossRef] [PubMed]

11. Joos, S.; Nettelbeck, D.M.; Reil-Held, A.; Engelmann, K.; Moosmann, A.; Eggert, A.; Hiddemann, W.; Krause, M.; Peters, C.; Schuler, M.; et al. German Cancer Consortium (DKTK)–A national consortium for translational cancer research. *Mol. Oncol.* **2019**, *13*, 535–542. [CrossRef]

12. Lawler, M.; Siu, L.L.; Rehm, H.L.; Chanock, S.J.; Alterovitz, G.; Burn, J.; Calvo, F.; Lacombe, D.; Teh, B.T.; North, K.N.; et al. All the world's a stage: Facilitating discovery science and improved cancer care through the global alliance for genomics and health. *Cancer Discov.* **2015**, *5*, 1133–1136. [CrossRef] [PubMed]

13. OECD. *Making Open Science a Reality*; OECD Science, Technology and Industry Policy Papers; OECD: Paris, France, 2015; Volume 25.

14. Besançon, L.; Peiffer-Smadja, N.; Segalas, C.; Jiang, H.; Masuzzo, P.; Smout, C.; Billy, E.; Deforet, M.; Leyrat, C. Open science saves lives: Lessons from the COVID-19 pandemic. *BMC Med. Res. Methodol.* **2021**, *21*, 117. [CrossRef]

15. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]

16. Sansone, S.A.; McQuilton, P.; Rocca-Serra, P.; Gonzalez-Beltran, A.; Izzo, M.; Lister, A.L.; Thurston, M. FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **2019**, *37*, 358–367. [CrossRef] [PubMed]

17. Centre, D.C. Disciplinary Metadata. Available online: https://www.dcc.ac.uk/guidance/standards/metadata (accessed on 12 December 2021).

18. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2019**, *47*, D23. [CrossRef]

19. Sarkans, U.; Füllgrabe, A.; Ali, A.; Athar, A.; Behrangi, E.; Diaz, N.; Fexova, S.; George, N.; Iqbal, H.; Kurri, S.; et al. From ArrayExpress to BioStudies. *Nucleic Acids Res.* **2021**, *49*, D1502–D1506. [CrossRef]

20. Madeira, F.; Park, Y.M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A.R.; Potter, S.C.; Finn, R.D.; et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47*, W636–W641. [CrossRef] [PubMed]

21. Burley, S.K.; Berman, H.M.; Kleywegt, G.J.; Markley, J.L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): The single global macromolecular structure archive. In *Protein Crystallography*; Humana Press: New York, NY, USA, 2017; pp. 627–641.

22. Pujar, S.; O'Leary, N.A.; Farrell, C.M.; Loveland, J.E.; Mudge, J.M.; Wallin, C.; Girón, C.G.; Diekhans, M.; Barnes, I.; Bennett, R.; et al. Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res.* **2018**, *46*, D221–D228. [CrossRef]

23. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Parkin, D.M.; Piñeros, M.; Znaor, A.; Bray, F. Cancer statistics for the year 2020: An overview. *Int. J. Cancer* **2021**, *149*, 778–789. [CrossRef]

24. Zhao, Z.; Zhang, K.N.; Wang, Q.; Li, G.; Zeng, F.; Zhang, Y.; Wu, F.; Chai, R.; Wang, Z.; Zhang, C.; et al. Chinese Glioma Genome Atlas (CGGA): A comprehensive resource with functional genomic data from Chinese gliomas. *Genom. Proteom. Bioinform.* **2021**, *19*, 1–12. [CrossRef]

25. Sarkans, U.; Chiu, W.; Collinson, L.; Darrow, M.C.; Ellenberg, J.; Grunwald, D.; Hériché, J.K.; Iudin, A.; Martins, G.G.; Meehan, T.; et al. REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nat. Methods* **2021**, *18*, 1418–1422. [CrossRef]

26. Touré, V.; Flobak, Å.; Niarakis, A.; Vercruysse, S.; Kuiper, M. The status of causality in biological databases: Data resources and data retrieval possibilities to support logical modeling. *Briefings Bioinform.* **2021**, *22*, bbaa390. [CrossRef]

27. Kingsley, J.L.; Costello, J.R.; Raghunand, N.; Rejniak, K.A. Bridging cell-scale simulations and radiologic images to explain short-time intratumoral oxygen fluctuations. *bioRxiv* **2021**, doi:10.1101/2021.03.18.435990. [CrossRef] [PubMed]

28. Hormuth, D.A.; Al Feghali, K.A.; Elliott, A.M.; Yankeelov, T.E.; Chung, C. Image-based personalization of computational models for predicting response of high-grade glioma to chemoradiation. *Sci. Rep.* **2021**, *11*, 8520. [CrossRef]

29. Jeanquartier, F.; Jean-Quartier, C.; Cemernek, D.; Holzinger, A. In silico modeling for tumor growth visualization. *BMC Syst. Biol.* **2016**, *10*, 59. [CrossRef] [PubMed]

30. Aerts, H.; Schirner, M.; Dhollander, T.; Jeurissen, B.; Achten, E.; Van Roost, D.; Ritter, P.; Marinazzo, D. Modeling brain dynamics after tumor resection using The Virtual Brain. *Neuroimage* **2020**, *213*, 116738. [CrossRef] [PubMed]

31. Bergmann, N.; Delbridge, C.; Gempt, J.; Feuchtinger, A.; Walch, A.; Schirmer, L.; Bunk, W.; Aschenbrenner, T.; Liesche-Starnecker, F.; Schlegel, J. The intratumoral heterogeneity reflects the intertumoral subtypes of glioblastoma multiforme: A regional immunohistochemistry analysis. *Front. Oncol.* **2020**, *10*, 494. [CrossRef] [PubMed]

32. Shi, Y.; Ding, D.; Liu, L.; Li, Z.; Zuo, L.; Zhou, L.; Du, Q.; Jing, Z.; Zhang, X.; Sun, Z. Integrative Analysis of Metabolomic and Transcriptomic Data Reveals Metabolic Alterations in Glioma Patients. *J. Proteome Res.* **2021**, *20*, 2206–2215. [CrossRef]

33. Yang, Y.; Sui, Y.; Xie, B.; Qu, H.; Fang, X. GliomaDB: A web server for integrating glioma omics data and interactive analysis. *Genom. Proteom. Bioinform.* **2019**, *17*, 465–471. [CrossRef]

34. Jean-Quartier, C.; Jeanquartier, F.; Holzinger, A. Open data for differential network analysis in glioma. *Int. J. Mol. Sci.* **2020**, *21*, 547. [CrossRef]

35. Jean-Quartier, C.; Jeanquartier, F.; Ridvan, A.; Kargl, M.; Mirza, T.; Stangl, T.; Markač, R.; Jurada, M.; Holzinger, A. Mutation-based clustering and classification analysis reveals distinctive age groups and age-related biomarkers for glioma. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 77. [CrossRef]

36. Jeanquartier, F.; Jean-Quartier, C.; Holzinger, A. Use case driven evaluation of open databases for pediatric cancer research. *BioData Min.* **2019**, *12*, 2. [CrossRef]

37. Ceccarelli, M.; Barthel, F.P.; Malta, T.M.; Sabedot, T.S.; Salama, S.R.; Murray, B.A.; Morozova, O.; Newton, Y.; Radenbaugh, A.; Pagnotta, S.M.; et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **2016**, *164*, 550–563. [CrossRef]

38. Krasnov, G.S.; Kudryavtseva, A.V.; Snezhkina, A.V.; Lakunina, V.A.; Beniaminov, A.D.; Melnikova, N.V.; Dmitriev, A.A. Pan-cancer analysis of TCGA data revealed promising reference genes for qPCR normalization. *Front. Genet.* **2019**, *10*, 97. [CrossRef]

39. Ortmayr, K.; Dubuis, S.; Zampieri, M. Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism. *Nat. Commun.* **2019**, *10*, 1841. [CrossRef] [PubMed]

40. Chang, Y.; Li, G.; Zhai, Y.; Huang, L.; Feng, Y.; Wang, D.; Zhang, W.; Hu, H. Redox regulator GLRX is associated with tumor immunity in glioma. *Front. Immunol.* **2020**, *11*, 3028. [CrossRef] [PubMed]

41. Feng, X.; Tustison, N.J.; Patel, S.H.; Meyer, C.H. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Front. Comput. Neurosci.* **2020**, *14*, 25. [CrossRef]

42. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* **2018**, arXiv:1811.02629.

43. Kofler, F.; Berger, C.; Waldmannstetter, D.; Lipkova, J.; Ezhov, I.; Tetteh, G.; Kirschke, J.; Zimmer, C.; Wiestler, B.; Menze, B.H. BraTS Toolkit: Translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. *Front. Neurosci.* **2020**, *14*, 125. [CrossRef]

44. Banerjee, S.; Mitra, S.; Masulli, F.; Rovetta, S. Glioma classification using deep radiomics. *SN Comput. Sci.* **2020**, *1*, 209. [CrossRef]

45. Lu, V.M.; Power, E.A.; Kerezoudis, P.; Daniels, D.J. The 100 most-cited articles about diffuse intrinsic pontine glioma: A bibliometric analysis. *Child's Nerv. Syst.* **2019**, *35*, 2339–2346. [CrossRef] [PubMed]

46. Akmal, M.; Hasnain, N.; Rehan, A.; Iqbal, U.; Hashmi, S.; Fatima, K.; Farooq, M.Z.; Khosa, F.; Siddiqi, J.; Khan, M.K. Glioblastome multiforme: A bibliometric analysis. *World Neurosurg.* **2020**, *136*, 270–282. [CrossRef] [PubMed]

47. Molinaro, A.M.; Taylor, J.W.; Wiencke, J.K.; Wrensch, M.R. Genetic and molecular epidemiology of adult diffuse glioma. *Nat. Rev. Neurol.* **2019**, *15*, 405–417. [CrossRef] [PubMed]

48. Celiku, O.; Johnson, S.; Zhao, S.; Camphausen, K.; Shankavaram, U. Visualizing molecular profiles of glioblastoma with GBM-BioDP. *PLoS ONE* **2014**, *9*, e101239. [CrossRef] [PubMed]

49. Thorsson, V.; Gibbs, D.L.; Brown, S.D.; Wolf, D.; Bortone, D.S.; Yang, T.H.O.; Porta-Pardo, E.; Gao, G.F.; Plaisier, C.L.; Eddy, J.A.; et al. The immune landscape of cancer. *Immunity* **2018**, *48*, 812–830. [CrossRef]

50. Prior, F.W.; Clark, K.; Commean, P.; Freymann, J.; Jaffe, C.; Kirby, J.; Moore, S.; Smith, K.; Tarbox, L.; Vendt, B.; et al. TCIA: An information resource to enable open science. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 1282–1285.

51. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82. [CrossRef] [PubMed]

52. Gao, J.; Mazor, T.; De Bruijn, I.; Abeshouse, A.; Baiceanu, D.; Erkoc, Z.; Gross, B.; Higgins, D.; Jagannathan, P.K.; Kalletla, K.; et al. The cBioPortal for Cancer Genomics. *Cancer Res.* **2021**, *81*, 207.

53. Pratt, D.; Chen, J.; Pillich, R.; Rynkov, V.; Gary, A.; Demchak, B.; Ideker, T. NDEx 2.0: A clearinghouse for research on cancer pathways. *Cancer Res.* **2017**, *77*, e58–e61. [CrossRef]

54. Pavlopoulou, A.; Spandidos, D.A.; Michalopoulos, I. Human cancer databases. *Oncol. Rep.* **2015**, *33*, 3–18. [CrossRef]

55. Avsec, Ž.; Kreuzhuber, R.; Israeli, J.; Xu, N.; Cheng, J.; Shrikumar, A.; Banerjee, A.; Kim, D.S.; Beier, T.; Urban, L.; et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **2019**, *37*, 592–600. [CrossRef]

56. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **2019**, *25*, 30–36. [CrossRef] [PubMed]

57. Matschinske, J.; Alcaraz, N.; Benis, A.; Golebiewski, M.; Grimm, D.G.; Heumos, L.; Kacprowski, T.; Lazareva, O.; List, M.; Louadi, Z.; et al. The AIMe registry for artificial intelligence in biomedical research. *Nat. Methods* **2021**, *18*, 1128–1131. [CrossRef] [PubMed]

58. Kleppe, A.; Skrede, O.J.; De Raedt, S.; Liestøl, K.; Kerr, D.J.; Danielsen, H.E. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **2021**, *21*, 199–211. [CrossRef]

59. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R.R.; et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **2020**, *10*, 1–12. [CrossRef] [PubMed]

60. Morid, M.A.; Borjali, A.; Del Fiol, G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput. Biol. Med.* **2020**, *28*, 104115. [CrossRef]

61. Commission, E. Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. 2021. Available online: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed on 12 December 2021).

62. Eberle, R.; Kaplan, D.; Montague, R. Hempel and Oppenheim on explanation. *Philos. Sci.* **1961**, *28*, 418–428. [CrossRef]

63. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]

64. Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.K.; Müller, K.R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer Nature: Cham, Switzerland, 2019; Volume 11700. [CrossRef]

65. Holzinger, A.; Malle, B.; Saranti, A.; Pfeifer, B. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Inf. Fusion* **2021**, *71*, 28–37. [CrossRef]

66. Holzinger, A.; Mueller, H. Toward Human-AI Interfaces to Support Explainability and Causability in Medical AI. *IEEE Comput.* **2021**, *54*, 78–86. [CrossRef]

67. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Mueller, H. Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef]

68. Holzinger, A.; Carrington, A.; Mueller, H. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. *KI-Kuenstliche Intell.* **2020**, *34*, 193–198. [CrossRef] [PubMed]

69. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.

70. Holzinger, A. Explainable ai and multi-modal causability in medicine. *i-com* **2020**, *19*, 171–179. [CrossRef]

71. Wulczyn, E.; Nagpal, K.; Symonds, M.; Moran, M.; Plass, M.; Reihs, R.; Nader, F.; Tan, F.; Cai, Y.; Brown, T.; et al. Predicting Prostate Cancer-Specific Mortality with AI-based Gleason Grading. *arXiv* **2020**, arXiv:2012.05197.

72. Das, T.; Andrieux, G.; Ahmed, M.; Chakraborty, S. Integration of online omics-data resources for cancer research. *Front. Genet.* **2020**, *11*, 578345. [CrossRef]

73. Wishart, D.S.; Mandal, R.; Stanislaus, A.; Ramirez-Gaona, M. Cancer metabolomics and the human metabolome database. *Metabolites* **2016**, *6*, 10. [CrossRef] [PubMed]

74. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef] [PubMed]

75. Zhao, H.; Heimberger, A.B.; Lu, Z.; Wu, X.; Hodges, T.R.; Song, R.; Shen, J. Metabolomics profiling in plasma samples from glioma patients correlates with tumor phenotypes. *Oncotarget* **2016**, *7*, 20486. [CrossRef] [PubMed]

76. Mayerhoefer, M.E.; Materka, A.; Langs, G.; Häggström, I.; Szczypiński, P.; Gibbs, P.; Cook, G. Introduction to radiomics. *J. Nucl. Med.* **2020**, *61*, 488–495. [CrossRef]

77. Diaz, O.; Kushibar, K.; Osuala, R.; Linardos, A.; Garrucho, L.; Igual, L.; Radeva, P.; Prior, F.; Gkontra, P.; Lekadir, K. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Phys. Medica* **2021**, *83*, 25–37. [CrossRef] [PubMed]

78. Shui, L.; Ren, H.; Yang, X.; Li, J.; Chen, Z.; Yi, C.; Zhu, H.; Shui, P. Era of radiogenomics in precision medicine: An emerging approach for prediction of the diagnosis, treatment and prognosis of tumors. *Front. Oncol.* **2020**, *10*, 3195.

79. National Academies of Sciences; Engineering, and Medicine. *Communicating Science Effectively: A Research Agenda*; National Academies Press: Washington, DC, USA, 2017.

80. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597.

81. Krogan, N.J.; Lippman, S.; Agard, D.A.; Ashworth, A.; Ideker, T. The cancer cell map initiative: Defining the hallmark networks of cancer. *Mol. Cell* **2015**, *58*, 690–698. [CrossRef] [PubMed]

82. Rodrigues, A.J.; Jin, M.C.; Wu, A.; Bhambhvani, H.P.; Li, G.; Grant, G.A. Risk of secondary neoplasms after external-beam radiation therapy treatment of pediatric low-grade gliomas: A SEER analysis, 1973–2015. *J. Neurosurgery Pediatr.* **2021**, *1*, 1–9. [CrossRef]

83. Kearney, A.; IQVIA. *Oncology Data Landscape in Europe Data Sources & Initiatives*; Technical Report; efpia: Brussels, Belgium, 2018.

84. Obermeyer, Z.; Emanuel, E.J. Predicting the future—big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **2016**, *375*, 1216. [CrossRef]

85. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing medical imaging data for machine learning. *Radiology* **2020**, *295*, 4–15. [CrossRef] [PubMed]

86. Marble, H.D.; Huang, R.; Dudgeon, S.N.; Lowe, A.; Herrmann, M.D.; Blakely, S.; Leavitt, M.O.; Isaacs, M.; Hanna, M.G.; Sharma, A.; et al. A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients. *J. Pathol. Inform.* **2020**, *11*, 22. [PubMed]

87. Cabitza, F.; Zeitoun, J.D. The proof of the pudding: In praise of a culture of real-world validation for medical artificial intelligence. *Ann. Transl. Med.* **2019**, *7*, 161. [CrossRef] [PubMed]

88. Zuiderwijk, A.; de Reuver, M. Why open government data initiatives fail to achieve their objectives: Categorizing and prioritizing barriers through a global survey. *Transform. Gov. People Process. Policy* **2021**, *15*, 377–395. [CrossRef]