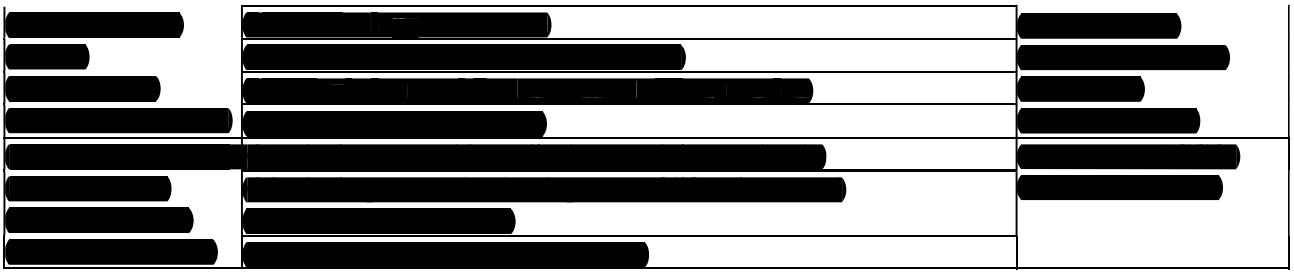
Spoke2 - Big Data-Open Data in Life Sciences

Table B3: Spoke 2, work packages subjects involved and associated tasks

WP	Tasks	Subjects involved
1. A holistic, innovative digital architecture for the storage and safe exchange of life sciences big data	1.1 A novel architecture for big data exchange	UNIMI, ALMAVIVA, TIM
	1.2 Governance, management and regulating the access to services of data in silico Sub-Tasks (pilots) • 1.2.1 Economic impact of a structured collection of big data in life sciences • 1.2.2 Pilot Exploiting Genomic Data for Drug Design • 1.2.3 Fusion of image-tabular data for federated learning of diagnostic models • 1.2.4 Pilot Digital Strategies for Active Ingredients Synthesis for Pharma • 1.2.5 Pilot Structure-based in silico target identification • 1.2.6 Green Radiotherapy	
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]



WP 1. A holistic, innovative digital architecture for the storage and safe exchange big data for life sciences

The goal of the WP is to conceive a novel digital infrastructure for the rapid exchange of data related to life sciences by means of machine learning and artificial intelligence, and test them in selected pilot cases.

Task 1.1 A novel architecture for big data exchange:
A platform will be created for the management of the whole lifecycle of life science data and information, fully respecting data protection regulations.

Task 1.2 Governance, management and regulating the access to services of data in silico:
This task deals with the governance, management, and regulatory actions related to the access to the services for elaborating the life science data, creating a marketplace of services based on the emerging needs of the research community. Services include i) interfaces for the platforms built for data collection ii) key nodes for collecting and giving value to the data flows from sensors iii) centralised or decentralised models to use and give value to data to calculate relevant inferences for life science. It will develop the strategy for data modeling, as well as for data protection and privacy.

PILOT CASES: The following tasks will be used to validate the solution proposed in Task 1.1 and Task 1.2.

Task 1.2.1 Economic impact of a structured collection of big data in life sciences:
The objective of this task is to conduct an assessment of the socio-economic impact of the research infrastructure for archiving and exchanging big data of life sciences developed within the Spoke 2 of the MUSA project. A platform for collecting and sharing big data in the life sciences sector can be understood, i.e. a structure that provides resources and services to scientific communities to conduct research and promote innovation in their fields (European Commission, 2017). As such, it is possible to measure its overall socio-economic impact using cost-benefit analysis, a consolidated approach for investment evaluation, and which has recently been used also for the evaluation of Research and Development, in particular related to large research infrastructures (Florio, 2019). The use of cost-benefit analysis for the social evaluation of research infrastructures requires considering their specificities in terms of generated benefits. These can be linked to the direct and indirect benefits obtained by the users of the infrastructure services: this is the case, for example, of the value for scientists of the publications produced thanks to the infrastructure, or of the positive spin-offs from which companies that contribute to the realization can take advantage of the same. These are accompanied by the benefits for the future use of knowledge (most often difficult to measure ex ante) and the intrinsic value of knowledge as a public good, i.e. the social value attributed to scientific discoveries by the community.

Task 1.2.2 Pilot Exploiting Genomic Data for Drug Design:
Creation of a virtual Biobank with data derived from RNA sequencing profiling of experimental humanized animal models. Validation of “secure transmission” strategy of the data generated

from the next generation sequencing platform and their digitalization. Application of artificial intelligence models for data analysis.

Task 1.2.3 Fusion of image-tabular data for federated learning of diagnostic models:
Creation of a repeatable approach for multi-centric diagnostic studies, based on federated learning and data of heterogeneous types. Preparation of local databases and standardization of metadata. Preparation and development of privacy-aware Artificial Intelligence models, characterized by high local throughput and exchange of parameters in federated learning mode. The reference case study focuses on techniques for merging multiple clinical information collected from independent centers for early diagnosis and staging of prostate cancer. The pilot will establish a repeatable AI pipeline that includes 1) collection of highly heterogeneous data (image type and tabular type) 2) digitization and fusion in order to (i) maximize the accuracy of early diagnosis (ii) assess its differential benefit compared to imaging alone by providing a reliable estimate of the gain in precision, compared to the cost of additional collection and pre-processing.

Task 1.2.4 Pilot Digital Strategies for Active Ingredients Synthesis for Pharma: Digital and technology-driven strategies for the development of innovative and sustainable synthesis of active ingredients and crucial intermediates, including antivirals, according to the green chemistry and circular economy principles.
Crucial elements and key technologies for the task will be: development of Flow chemistry for ML-controlled accelerated reactions; design, implementation and optimization of organic transformation under continuous flow conditions; development of an automated research line for organic chemistry synthesis of products of interest for high value pharma drug design.

Task 1.2.5 Pilot Structure-based in silico target identification
The task involves the development of a structure-based method for target identification by exploiting optimized approaches of inverse docking simulations. Thus, the task aims to identify which proteins of potential medicinal role are targeted by a given compound.
The developed workflow will be implemented into a web service to allow its user-friendly and distributed exploitation. The in silico identification of the potential therapeutic targets for a given ligand can find many relevant applications such as for the interpretation of the phenotyping screening and for the mechanistic rationalization of the effects of natural compounds.

Task 1.2.6 Green Radiotherapy:
Healthcare has a large carbon footprint, estimated around the 5% of the entire carbon emissions in Europe. In this scenario is a call to action to provide more sustainable oncology practices and reduce its carbon footprint. Since more than half of cancer patients receive some form of radiation therapy (RT) throughout the course of the disease and about 25% are going to be irradiated more than once during their oncological history, there is great interest in understanding the potential environmental impact of RT. Aim of the present project is to develop a carbon footprint score to evaluate CO2 emissions produced due to RT treatments in one of the largest RT centers in Lombardy and Italy. Then score is then going to o be applied to other radiation oncology centers across Lombardy and Italy exploiting the MUSA infrastructure. The score is going to incorporate contributions from treatment commute, pre-treatment imaging and radiation delivery/fractionation (fx) in patients treated between 2012 and 2022 at the Radiation Oncology Department of IEO, European Institute of Oncology IRCCS. A particular focus is going to be dedicated to assessing how the evolution of treatment technology over the past 10 years has impacted carbon footprint, cost, and sustainability. Specifically, it will be estimated whether the move towards more hypofractionated treatments (fewer and higher - >2 Gy – doses), in particular during Covid pandemics, has significantly reduced the carbon footprint of RT treatments along with a reduction of the amount of patient travel.
The results coming from this analysis will likely be able to provide a focus for considerations of where initial reductions in carbon footprint could be aimed, to motivate further studies and to shed a light on the topic in general. Wide adoption of these approaches can substantially increase the

sustainability of cancer treatment and, considering the world-wide scale of cancer, might have an immediate impact of the world ecosystem.

[redacted]