machine learning & knowledge extraction

Viewpoint

A Machine Learning Perspective on Personalized Medicine: An Automized, Comprehensive Knowledge Base with Ontology for Pattern Recognition

Frank Emmert-Streib [1,2,*] **and Matthias Dehmer** [3,4,5]

[1] Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology, Tampere 33720, Finland
[2] Institute of Biosciences and Medical Technology, Tampere 33520, Finland
[3] Department of Mechatronics and Biomedical Computer Science, UMIT, Hall in Tyrol 6060, Austria; matthias.dehmer@umit.at
[4] College of Computer and Control Engineering, Nankai University, Tianjin 300071, China
[5] Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Steyr Campus 4400, Austria
* Correspondence: v@bio-complexity.com; Tel.: +358-50-301-5353

Received: 28 July 2018; Accepted: 4 September 2018; Published: 8 September 2018

Abstract: Personalized or precision medicine is a new paradigm that holds great promise for individualized patient diagnosis, treatment, and care. However, personalized medicine has only been described on an informal level rather than through rigorous practical guidelines and statistical protocols that would allow its robust practical realization for implementation in day-to-day clinical practice. In this paper, we discuss three key factors, which we consider dimensions that effect the experimental design for personalized medicine: (I) phenotype categories; (II) population size; and (III) statistical analysis. This formalization allows us to define personalized medicine from a machine learning perspective, as an automized, comprehensive knowledge base with an ontology that performs pattern recognition of patient profiles.

Keywords: machine learning; pattern recognition; personalized medicine; precision medicine; genomics

1. Introduction

Personalized medicine, or precision medicine [1], is a new paradigm that gained considerable interest within the last few years in the medical and biomedical research community [2–6]. Initially, sparked by the technologies that have been developed for the *Human Genome Project* [7–9] there is an increasing effort to convert the principle ideas underlying personalized medicine into the medical and clinical practice. An informal definition of personalized medicine has been provided by Ginsburg et al. [10] stating:

> Personalized medicine is a broad and rapidly advancing field of health care that is informed by each person's unique clinical, genetic, genomic, and environmental information.

So far, numerous articles have been published elaborating on different aspects around the above informal definition [2,11–13]. However, currently, there exist no practical design or implementation protocols that would provide a well-defined realization of personalized medicine [14]. As a consequence, the experimental design suitable to establish the paradigm of personalized medicine as a clinical standard, is largely unclear.

In this paper, we elaborate on three key factors that are in our opinion crucial in order to enable a practical realization of personalized medicine. These factors are also important for the definition of its experimental design. As a result from our discussion, we will provide a definition of personalized medicine from a machine learning perspective. This definition will be more clear for the machine learning community than the above given informal characterization by Ginsburg [10].

2. Three Key Factors of Personalized Medicine

There are many factors that have to be taken into account in the experimental design of a biomedical study. In the following framework of personalized medicine, we highlight three major factors: (I) phenotype categories; (II) population size; and (III) statistical analysis. In Figure 1A, we illustrate these three factors as dimensions, whereas for each dimension we distinguished between the ideal setting (unlimited population size for instance) and a realistic scenario (limited population size with under-representation of some disease phenotypes for instance).

The first factor, referred to as "phenotype categories", represents our knowledge about human diseases, including their potential subtypes. For instance, the evolving definition of breast cancer molecular subtypes [15–17] is an example for our incomplete knowledge of complex diseases such as cancer. In their seminal work, Perou and colleagues identified up to five molecular subtypes [15,18,19]. More recently, Curtis et al. jointly analyzed copy number alterations and gene expression profiles from the largest breast cancer dataset to date (2000 tumors; referred to as METABRIC) and discovered 10 different subtypes [17]; however, the robustness of this new classification still remains to be demonstrated. The problem that results from this incomplete recognition of disease phenotypes is that a disorder that is not known *cannot* be screened and investigated. This is illustrated in Figure 1 by the fainted green patients at the beginning of the dimension "phenotype categories" reflecting our real knowledge.

In a similar way, the lack of differentiating between subtypes of diseases may impair the treatment of diseases because every patient within the broader categories will receive the same treatment despite the fact that a different medication or treatment may be more appropriate. For instance, for breast cancer, chemotherapy is predominantly used in estrogen receptor-negative (ER-) subtype. In contrast, hormone therapy, e.g., using tamoxifen or anastrozole, is used for treating estrogen receptor (ER+) and progesterone receptor (PR+) subtypes.

The second factor, referred to as "population size", is important to consider as a separate dimension because, for orphan diseases, which effect only a very small fraction of the human population, it is practically not possible to collect data from patients with an arbitrarily large sample size. For instance, for ribose-5-phosphate isomerase deficiency [20], there is so far only a single patient diagnosed [21]. Interestingly, the number of such disorders is surprisingly large. According to [22], there are more than 5000 disorders categorized as rare. These examples emphasize an important limitation that needs to be considered appropriately.

It is important to emphasize that this factor is not independent of the first factor, the phenotype category. The reason for this is that it is not sufficient to enrol a certain number of patients from breast cancer, but these numbers need to cover all known subtypes in a homogeneous way.

The third factor, referred to as "statistical analysis", represents a statistical inference mechanism to classify, predict, or diagnose a patient by means of statistical methods. In this setting, machine learning approaches, such as pattern recognition, are widely used to develop statistical models from a dataset in order to obtain quantitative results that are subsequently used for inference. However, the quality of a statistical data analysis depends on a variety of factors, including sample size, effect size, and the applied method itself leading to, in reality, an imperfect performance. For instance, for every statistical hypothesis test, one needs to quantify the significance level α, which corresponds to the type I error we are willing to making indicating the false positive probability [23].

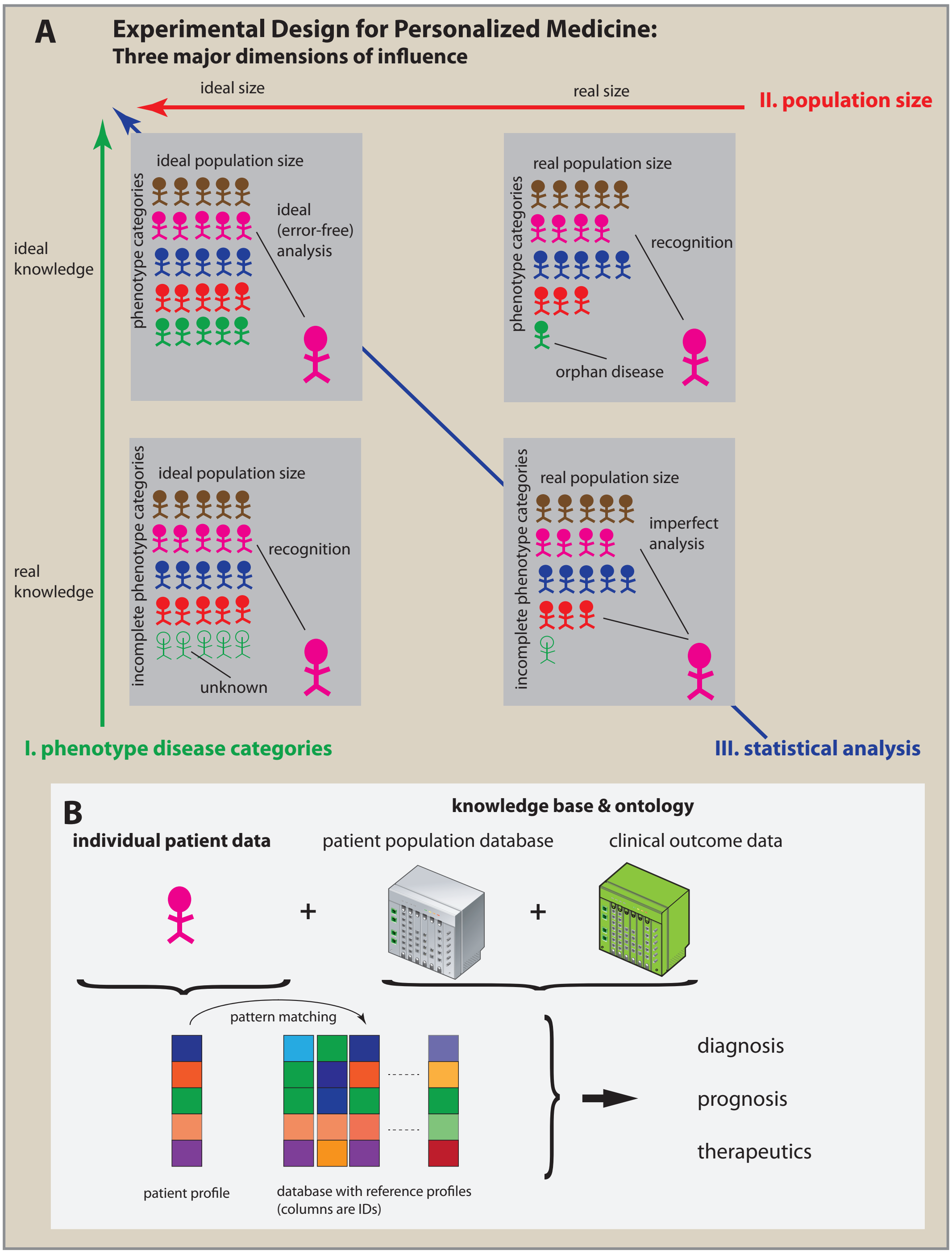
A
Experimental Design for Personalized Medicine:
Three major dimensions of influence
ideal size
real size
II. population size
ideal knowledge
real knowledge
ideal population size
phenotype categories
ideal (error-free) analysis
real population size
recognition
orphan disease
incomplete phenotype categories
unknown
imperfect analysis
I. phenotype disease categories
III. statistical analysis
B
knowledge base & ontology
individual patient data
patient population database
clinical outcome data
+
+
pattern matching
patient profile
database with reference profiles (columns are IDs)
diagnosis
prognosis
therapeutics

Figure 1. (**A**) Overview of three major factors of influence on personalized medicine. Together these factors define the experimental design of the field. (**B**) Simplified working mechanism of the automized comprehensive knowledge base with an ontology that performs pattern recognition of patient profiles.

3. Advances Required to Implement Personalized Medicine

Ideally, we would like to have complete knowledge about all existing phenotype categories, including disorders, and for each of these phenotype categories, we would have a population of patients of unlimited size. In this setting, the statistical method we apply would allow ideally an error free analysis. This ideal state is represented in the top left corner in Figure 1A if one follows the three dimensions toward their arrow heads. Practically, this ideal state is of course unachievable due to realistic limitations; however, advances in the three key dimensions of personalized medicine will result in an *approximation* of such an ideal state.

A few remarks to improve upon these three factors:

Population size: Enrolling more patients can improve visibility and facilitate the organization of large clinical trials (community-based/open consortia to allow anybody to contribute; strict regulation still required to avoid garbage-in/garbage-out). A particular problem in this respect is posed by orphan diseases because these would not allow the recruitment of large populations to ensure a sound statistical analysis. Importantly, personalized medicine does not provide solutions for this problem but faces the same problems as traditional medicine.

Phenotype categories: Biomedicine is an intense research field with many laboratories competing and collaborating to better understand biological mechanisms underlying human diseases and their molecular characteristics (genotypes) and phenotypes. Better phenotype categorization can be achieved by better sharing of published and unpublished data/code and results (ontologies, MeSH terms, etc.) as well as more efficient use of biotechnologies and more collaboration to exchange expertise, expensive, and/or cutting-edge technologies through large consortia.

Aside from these specific factors, it is important to emphasize that the research required to obtain a better understanding of phenotype categories is *basic* (or fundamental) and not translational. This is important to note because, currently, translational medicine is a hot field, but this is a reminder that one can only translate the knowledge to the patient that is actually available [24]. Hence, the lack of the identification of disease categories leads to a lack of the identifiability on the clinical side, with negative consequences for diagnosis, treatment, and prognosis for the patient. For this reason, funding needs to be carefully balanced between translational and basic biomedical research in order to avoid a negative bias.

Statistical analysis: The performance of statistical methods for pattern recognition mainly rely on the prior knowledge with respect to the disease and problem under investigation, quantity, and quality of data, the type of data (the technology used must be relevant to the problem that needs to be addressed), and the algorithm itself. Missing data, for instance, are a factor that can considerably lower the quality of the data despite the fact that a sufficiently large sample may be available. Since missing data are unavoidable in any study (their amount can only be controlled by a good data management but not eliminated), this demonstrates also the general importance of statistical methods.

Statistical analysis is likely the area where the advances might be the most difficult to implement for two reasons. First, computational genomics or similar areas are relatively young research fields where major improvements are still required to obtain a more efficient analysis. Specifically, due to the uniqueness of the data characteristics of new technologies, e.g., next-generation sequencing, the application of proprietary software seems inappropriate at times because such packages are not at the forefront of the statistical analysis developments and are falling behind. This is for instance due to the need for the time-consuming establishment of graphical-user-interfaces (GUIs) in order to make the methods accessible for "non-computational biologists". For this reason, open software endeavours such as R and Bioconductor [25,26] should receive stronger support because they enable the most efficient adaptation of statistical methods to high-dimensional genomics data. As a positive side-effect, R comes without license fees and, hence, allows the free distribution of software packages within the community to make analysis results fully reproducible [27,28]. This is in contrast to analysis results that have been obtained by using proprietary software because if one does not have a license, one cannot reproduce the results.

Second, many clinicians and biologists still do not consider statistical analysis, e.g., in the form of computational genomics, as a coequal contribution to biomedicine or biology. For this reason, the education in this area is taken too lightly. This is evidenced by the lack of understanding of many physicians, house officers, and students to correctly interpret the results of diagnostic tests, as found in [29]. Unfortunately, this has not changed over the last three decades [30], indicating the severity of this problem. This implies an urgent need for deeper educational changes in the curriculum of students in the life and clinical science in the understanding and usage of statistical analysis methods.

4. A Machine Learning Perspective

Based on the above discussion of personalized medicine, we are now able to formally characterize this field from a machine learning perspective. From this perspective, we can state:

> Personalized medicine is an automized comprehensive knowledge base with an ontology that performs pattern recognition of patient profiles.

We visualize a simplified working mechanism of this in Figure 1B. For a patient who presents at a clinic, a variety of quantitative clinical and genomic measurements are conducted, leading to the creation of a patient's profile vector. Each component of this vector corresponds to a biomarker (or feature) necessary to perform statistical analysis. For instance, a profile vector can correspond to gene expression values measured either by DNA microarrays or next-generation sequencing technologies [31–33]. In this case, a component of a vector represents the gene expression value of an individual gene, mRNA, or even non-coding RNA if RNA-seq data are available. Interestingly, such profile vectors are not sparse but contain informative measurements about the molecular activity of a phenotype on a genomic scale. In contrast, inferred regulatory networks from such data are usually sparsely connected [34].

In simple terms, such an analysis consists in the comparison/matching of the profile vector of the patient with a comprehensive knowledge base that provides similar profile vectors for a population of patients together with clinical outcome variables, e.g., survival time or reaction to medications. The connection between reference profiles and clinical outcome variables establishes an ontology [35,36] that allows for inferences for the profile of the patient. Practically, this means that the *best matching* of the profiles of a patient allows for recommendations (representing the inference step) for the diagnosis, prognosis, and therapeutics of the patient. Ideally, these recommendations are actual decisions in an automized procedure. This automatization step would also be crucial for distinguishing personalized medicine from traditional medicine, which can be seen as a doctor-in-the-loop [37].

The comprehensiveness of the knowledge base and the ontology is important because, as discussed above for the "phenotype categories", a lack of reference data leads to the omission of disease phenotypes. This implies that personalized medicine is inherently a community effort because individual groups cannot cover the entirety of medicine but can only contribute incremental information.

Finally, we want to note that the three key factors discussed above have a direct effect on the profiles of the patient as well as the knowledge base and the ontology in an interconnected manner. Hence, the quality of the implemented personalized medicine solution is crucially affected by these three factors.

5. Practical Personalized Medicine

The above given formal characterization of personalized medicine from a machine learning perspective can be seen as idealistic. The reason for this is twofold. First, the comprehensiveness of the knowledge base as well as the oncology assume a complete understanding of all disorders including their manyfold subtypes as well as known treatments for each of these. Currently, we are far away from this stage and in fact have severely incomplete knowledge. Nevertheless, the resulting knowledge base and oncology that can be constructed based on our current information can be very useful. The second reason is with respect to the automatization of the whole process. The realization of this step, at least in a comprehensive manner, is also currently out of reach.

Removing both idealistic requirements from the above characterization, we reach the following pragmatic working definition of personalized medicine.

> Working definition: Personalized medicine is a knowledge base with an ontology that performs pattern recognition of patient profiles.

Formulated in this way, it provides a clearer view on the essential four components of this characterization which are as follows:

1. knowledge base;
2. ontology;
3. pattern recognition;
4. patient profiles.

From a more theoretical perspective, the knowledge base can be considered as *prior information*, and the patient profiles as data. Hence, both components refer to different kinds of "data" used and needed for an analysis. The oncology component, on the other hand, adds the semantic meaning to these two kinds of data by connecting both with each other and clinically relevant interpretations. Finally, the pattern recognition component is the working horse that deals with the various uncertainties and errors within the two kinds of data.

6. Closing the Loop

The above dissection of our characterization of personalized medicine into a working definition allows the back-connection to our discussion on the experimental design for personalized medicine (see also Figure 1A). Specifically, the ontology component of our characterization corresponds to the *phenotype disease categories*, whereas the data components of the characterization, the knowledge base and the patient profiles, can be identified with the data dimension we termed *population size*. Lastly, the pattern recognition component corresponds to the *statistical analysis* dimension.

By relating the characterization of personalized medicine with its experimental design, we have a direct approach to its practical realization. This enables without idealistic assumptions a case-by-case implementation of personalized medicine, e.g., for breast cancer and its various subtypes. In this way, a growing knowledge base and ontology for different disorders over time can be eventually integrated with each other to become comprehensive.

7. Conclusions

The introduction of any new field creates excitement about its potential application areas but also confusion with respect to its practical realization. This is in particular the case for personalized and precision medicine because the wealth of data that accompany these fields poses considerable challenges for their analysis [38]. We think that our discussion describing personalized medicine as an automized comprehensive knowledge base with an ontology that performs pattern recognition of patient profiles eliminates much of this confusion by providing a clear formal definition of this field. This would demystify personalized medicine for the machine learning community so that the focus can be placed on its realization.

Author Contributions: F.E.-S. conceived the study. All authors contributed to the writing of the manuscript and approved the final version.

Funding: M.D. thanks the Austrian Science Funds for supporting this work (project P30031).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Katsnelson, A. Momentum grows to make 'personalized' medicine more 'precise'. *Nat. Med.* **2013**, *19*, 249. [CrossRef] [PubMed]
2. Auffray, C.; Chen, Z.; Hood, L. Systems medicine: The future of medical genomics and healthcare. *Genome Med.* **2009**, *1*, 2. [CrossRef] [PubMed]
3. Chin, L.; Andersen, J.N.; Futreal, P.A. Cancer genomics: From discovery science to personalized medicine. *Nat. Med.* **2011**, *17*, 297–303. [CrossRef] [PubMed]
4. Chen, R.; Snyder, M. Promise of personalized omics to precision medicine. *Wiley Interdiscip. Rev.* **2013**, *5*, 73–82. [CrossRef] [PubMed]
5. Seo, D.; Ginsburg, G.S. Genomic medicine: Bringing biomarkers to clinical medicine. *Curr. Opin. Chem. Biol.* **2005**, *9*, 381–386. [CrossRef] [PubMed]
6. Tian, Q.; Price, N.D.; Hood, L. Systems cancer medicine: Towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J. Intern. Med.* **2012**, *271*, 111–121. [CrossRef] [PubMed]
7. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431*, 931–945. [CrossRef] [PubMed]
8. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [PubMed]
9. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304–1351. [CrossRef] [PubMed]
10. Ginsburg, G.S.; Willard, H.F. Genomic and personalized medicine: Foundations and applications. *Transl. Res.* **2009**, *154*, 277–287. [CrossRef] [PubMed]
11. Emmert-Streib, F.; Tuomisto, L.; Yli-Harja, O. The Need for Formally Defining 'Modern Medicine' by Means of Experimental Design. *Front. Genet.* **2016**, *7*, 60. [CrossRef] [PubMed]
12. Gonzalez-Angulo, A.M.; Hennessy, B.T.; Mills, G.B. Future of Personalized Medicine in Oncology: A Systems Biology Approach. *J. Clin. Oncol.* **2010**, *28*, 2777–2783. [CrossRef] [PubMed]
13. Welch, B.M.; Kawamoto, K. Clinical decision support for genetically guided personalized medicine: A systematic review. *J. Am. Med. Inform. Assoc.* **2012**, *20*, 388–400. [CrossRef] [PubMed]
14. Lesko, L.J. Personalized medicine: Elusive dream or imminent reality? *Clin. Pharmacol. Ther.* **2007**, *81*, 807–816. [CrossRef] [PubMed]
15. Sorlie, T.; Tibshirani, R.; Parker, J.; Hastie, T.; Marron, J.S.; Nobel, A.; Deng, S.; Johnsen, H.; Pesich, R.; Geisler, S.; et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 8418–8423. [CrossRef] [PubMed]
16. Perou, C.M.; Sorlie, T.; Eisen, M.B.; Van De Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; et al. Molecular portraits of human breast tumours. *Nature* **2000**, *406*, 747–752. [CrossRef] [PubMed]
17. Curtis, C.; Shah, S.P.; Chin, S.F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **2012**, *486*, 346. [CrossRef] [PubMed]
18. Prat, A.; Parker, J.; Karginova, O.; Fan, C.; Livasy, C.; Herschkowitz, J.; He, X.; Perou, C. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **2010**, *12*, R68. [CrossRef] [PubMed]
19. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **2008**, *455*, 1061–1068. [CrossRef] [PubMed]
20. Huck, J.H.; Verhoeven, N.M.; Struys, E.A.; Salomons, G.S.; Jakobs, C.; van der Knaap, M.S. Ribose-5-phosphate isomerase deficiency: New inborn error in the pentose phosphate pathway associated with a slowly progressive leukoencephalopathy. *Am. J. Hum. Genet.* **2004**, *74*, 745–751. [CrossRef] [PubMed]
21. Wamelink, M.M.; Grüning, N.M.; Jansen, E.E.; Bluemlein, K.; Lehrach, H.; Jakobs, C.; Ralser, M. The difference between rare and exceptionally rare: Molecular characterization of ribose 5-phosphate isomerase deficiency. *J. Mol. Med.* **2010**, *88*, 931–939. [CrossRef] [PubMed]

22. Schieppati, A.; Henter, J.I.; Daina, E.; Aperia, A. Why rare diseases are an important medical and social issue. *Lancet* **2008**, *371*, 2039–2041. [CrossRef]
23. Lehman, E. *Testing Statistical Hypotheses*; Springer: Berlin, Germany, 2005.
24. Mankoff, S.P.; Brander, C.; Ferrone, S.; Marincola, F.M. Lost in translation: Obstacles to translational medicine. *J. Transl. Med.* **2004**, *2*, 14. [CrossRef] [PubMed]
25. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0.
26. Gentleman, R.; Carey, V.; Bates, D.E.A. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80. [CrossRef] [PubMed]
27. Diggle, P.J.; Zeger, S.L. Embracing the concept of reproducible research. *Biostatistics* **2010**, *11*, 375. [CrossRef] [PubMed]
28. Peng, R.D. Reproducible Research in Computational Science. *Science* **2011**, *334*, 1226–1227. [CrossRef] [PubMed]
29. Casscells, W.; Schoenberger, A.; Graboys, T.B. Interpretation by Physicians of Clinical Laboratory Results. *N. Engl. J. Med.* **1978**, *299*, 999–1001. [CrossRef] [PubMed]
30. Manrai, A.; Bhatia, G.; Strymish, J.; Kohane, I.; Jain, S. Medicine's uncomfortable relationship with math: Calculating positive predictive value. *JAMA Intern. Med.* **2014**, *174*, 991–993. [CrossRef] [PubMed]
31. Metzker, M. Sequencing technologies-the next generation. *Nat. Rev. Genet.* **2009**, *11*, 31–46. [CrossRef] [PubMed]
32. Stupnikov, A.; Tripathi, S.; de Matos Simoes, R.; McArt, D.; Salto-Tellez, M.; Glazko, G.; Emmert-Streib, F. samExploreR: Exploring reproducibility and robustness of RNA-seq results based on SAM files. *Bioinformatics* **2016**, *32*, 3345–3347. [CrossRef] [PubMed]
33. Quackenbush, J. Microarray analysis and tumor classification. *N. Engl. J. Med.* **2006**, *345*, 2463–2472. [CrossRef] [PubMed]
34. Emmert-Streib, F.; de Matos Simoes, R.; Glazko, G.; McDade, S.; Haibe-Kains, B.; Holzinger, A.; Dehmer, M.; Campbell, F. Functional and genetic analysis of the colon cancer network. *BMC Bioinform.* **2014**, *15*, 6. [CrossRef] [PubMed]
35. Chandrasekaran, B.; Josephson, J.R.; Benjamins, V.R. What are ontologies, and why do we need them? *IEEE Intell. Syst. Their Appl.* **1999**, *14*, 20–26. [CrossRef]
36. Fonseca, F. The double role of ontologies in information science research. *J. Am. Soc. Inform. Sci. Technol.* **2007**, *58*, 786–793. [CrossRef]
37. Kieseberg, P.; Malle, B.; Frühwirt, P.; Weippl, E.; Holzinger, A. A tamper-proof audit and control system for the doctor in the loop. *Brain Inform.* **2016**, *3*, 269–279. [CrossRef] [PubMed]
38. Holzinger, A.; Dehmer, M.; Jurisica, I. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC Bioinform.* **2014**, *15*, I1. [CrossRef] [PubMed]