medidata | 2017 年 5 月白皮书

利用大数据挖掘，充分发挥您的临床试验数据和内容的价值

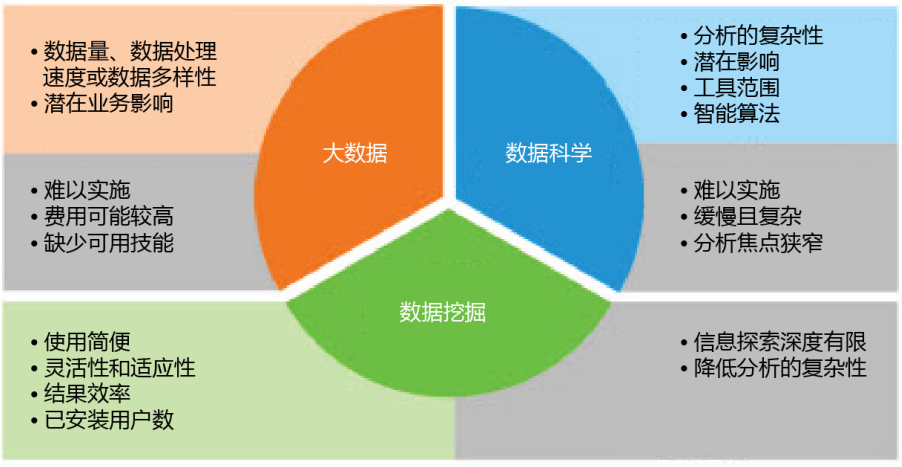
大数据 (Big Data) 重新定义了我们的世界

从休闲娱乐，到我们自身的健康，再到我们工作的行业 - 大数据已经改变了我们的世界。像 Netflix 和 Amazon Prime 这样的订阅式内容的提供商正进行电视节目的转型，他们利用详细的观众分类和观影习惯数据来重新思考新节目该如何备资、制作和投放市场。每天，像 Nest 这样的装置都在通过收集整合传感器数据，使家庭供暖和降温逐步实现由自动恒温器控制。

在临床研究行业，更多数据点的建立及其有效性也在改变一切。以前，数据都来自单一来源：患者去诊所看病，然后该患者的信息会被录入一个电子数据采集（EDC）系统。如今，临床试验可以容纳数量惊人的多种形式数据和内容：从传统临床数据，到高分辨率图像，再到基因组和可穿戴式传感器数据、研究者文件、知情同意书以及其他更多形式。这样的数据爆炸带来了崭新的变革性机遇，但同时也伴随着额外的风险。

大数据并不是全部

根据 Gartner 的定义，大数据仅仅是三元模型中的一个元素：大数据、数据挖掘 (Data Discovery) 和数据科学 (Data Science) 三者的结合创建了一个模型，Gartner 将之定义为**大数据挖掘 (Big Data Discovery)**。在临床试验中，大数据挖掘主要用于在临床研究过程中创建和使用交互报告，以及探索来自多个来源的数据（例如成像和移动医疗数据），从而得出可行的结论。最后，模型的这三个元素要求简化工具（以适用广大用户的需求），并访问标准化和多种类数据源。



来源：ZDNet.com & Gartner。大数据挖掘将大数据、数据科学以及数据挖掘相结合。

1. 大数据

如今，技术行业在描述大数据时通常使用五个关键特征：多样性、体量、速度、真实性和价值性。临床试验目前使用**多源化**数据采集，数据量远远超过传统的诊所内采集的数据量。每个来源在补充临床研究的有效性方面带来了独一无二的机会，同时也存在不同于其他来源的挑战。例如，医学影像资料、X 光片、CT 扫描和 MRI 为患者的总体情况提供更深入的了解，但它们通常与采集传统临床数据的系统分开存储和管理。

其次，临床研究人员必须考虑大数据的巨大**体量**。许多新型的数据形式本身具有大体量特征。一份 MRI 数据就可以有上千兆字节大小。因此搭架一个支持从业者上传、下载、更改和查询大量数据的 IT 架构将会成为挑战。

数据速度与数据采集频率有关。举例来说，传感器数据每秒钟为研究人员提供每个病人的大量的测量结果。

数据多样性、体量和速度不可避免的将我们引向**数据真实性**问题：我们该如何保证我们临床试验中数据的准确性和完整性？本质上讲，从多种来源采集到的数据必须经过清理、标准化和验证。

最后，结合数据的多样性、体量、速度和真实性创建而成的强大、标准化数据集能够强化分析力和洞察力，如果在进行临床试验时正确加以利用，那么这些数据最终会发挥巨大的**价值**。临床研究人员、申办方和 CRO 等类似的组织都可以利用数据和内容来优化流程的各个方面。仅凭采集大数据并进行标准化而不涉及数据挖掘和数据科学，不能充分获得深刻见解，最终推动实现临床开发过程。

2. 数据挖掘

数据挖掘工具（如仪表板和基准测试程序）使数据分析更加简便、快捷、灵活，从而方便广大不需要具备数据科学或统计建模方面专业技能的使用者进行访问。人们把这些新的使用者称为市民数据科学家 (Citizen Data Scientist)。

在临床试验中，数据挖掘至关重要。这并不是需要分析全部数据，而是显现相关数据来支持具体的决策，例如患者参与、研究中心可行性或研究中心的监测。确保交叉研究报告和针对具体研究参数的专用报告能够支持大量的研究微调，从而对可行性或效能产生重大影响。

为了充分发挥大数据的价值，临床开发技术必须不仅能提供对大量多样化数据进行快速采集、标准化处理和验证的功能，还要能够提供易于使用的报告和仪表板功能。这样做使得从事临床试验的庞大生态体系涉及数据、监测、操作或监管（工作）能够优化其日常活动的各个流程。

3. 数据科学

虽然数据挖掘可以加强临床开发生态体系中的市民数据科学 (Citizen Data Science)，但是数据科学及其提供的分析功能将为临床开发转型提供许多机会。这些包括研究中心选择 (Site Selection)、基于风险的监察 (Risk Based Monitoring)，甚至是拟真对照组 (Synthetic Control Arm) 等方面的新进展。更重要的是，数据科学对临床试验的改进程度与其数据集的稳健性成正比。与此类似的一个概念是群体智慧。当生命科学行业拥有整合数以万计的试验、治疗领域、地域和申办方的数据并能够进行标准化的解决方案时，即可得到内容更为丰富的数据集。

将这些庞大的数据集与新型数据科学技术（例如人工智能和机器学习）进行组合，可以阐明临床试验因素之间的未知关系。在临床试验操作中，数据科学工具的应用不仅在仪表板和基准测试程序方面支持数据挖掘的进行，而且还会直接影响试验的入组、研究中心管理、不良事件预测和药品供应。

例如，机器学习算法可以从历史临床试验数据中进行学习，以便实现不良事件自有文本的自动标准化。这个标准化可以改善试验中的信号检测，从而更早地识别出安全风险。同样地，算法也可以被训练成通过观察相关临床特征来检测此类事件是否属于严重级别。这种自动分类可以缩短向监管机构报告的时间，是临床试验的关键组成部分之一。为了真正发挥数据科学在临床开发中的作用，这些工具需要从能够采集、标准化和汇总数据的灵活平台中获得，在增加智能化的同时不影响数据的真实性。

在临床试验操作中充分发挥大数据挖掘的价值

电子数据采集系统（EDC）掌握了大数据的关键内容，但它不会自行释放数据挖掘和数据科学的潜力。由于数据的来源不同，采用电子数据采集系统进行整合是一种非常有效的方法。临床开发中所用的最先进 EDC 已不仅仅简单采集传统的医院内数据，而是整合多个来源的数据，同时对数据进行清理、标准化和验证。

EDC 的核心是为采集的所有数据提供单一的数据存储区（均储存在一处），从而成为研究的单一数据源。然后利用其单一数据集推动数据挖掘和数据科学，进而实时报告、分析和作出研究实施决策。

最后，EDC 可打下基础。接下来是将数据挖掘和数据科学技术以及执行这些技术所需的人力资本交予临床开发从业者，由他们根据这些采取行动。这通常是临床开发组织及其重要的 IT 合作伙伴的关键决策点。

除了有效管理当前正在进行的研究外，人们是否还希望为数据分析和数据挖掘提供技术？或者，他们是否应直接利用已经植入以上技术的成熟解决方案？

现今，最先进的临床开发技术提供者会提供数据挖掘和数据科学技能。利用这些技能可以减轻申办方和 CRO 的未来开发负担。本文的其余部分将要探讨的是，利用这些技能还可用于未来的临床研究。

下文的三个使用案例说明了如何在临床试验过程中应用大数据、数据挖掘和数据科学。

在适应性临床试验中利用大数据

按照适应性临床试验方案的规定审查数据并对研究设计实时变更，这一能力能够为适应性临床试验带来新突破。实时研究设计变更可让申办方和合作伙伴在几分钟内更有效且高效地扩大和缩小规模。借助这些类型的技能，所有申办方很快就能够轻松执行高度复杂的试验设计。

适应性临床试验将会有何优势？

适应性临床试验能预先计划，因此可根据中期分析修改研究设计和假设。在研究内的计划时间点以全盲或非盲进行累积研究数据的分析，并且可在无正式统计假设检测的情况下进行。

与非适应性研究相比，适应性设计方法可实现：更有效地提供相同信息的研究，增加研究目的取得成功的可能性的研究，或产生对治疗效果的理解性提高的研究（例如，对剂量反应关系或亚组效应进行更好地估计也可能实现更有效的后续研究）。另外，适应性临床试验设计可基于患者安全性的剂量反应进行滴定，这在高风险疾病领域尤其重要。

有哪些问题？

需要进行前期计划和策略思考，从而确保适应性临床试验取得成功。在一项研究中结合试验设计可推动更深入的分析。然而，这需要在准备阶段进行周密的计划，并需要相当大的投资，且在每次做出改变时需要延长数周。在某些情况下，适应性临床试验因其复杂性而被完全忽视。因为适应性临床试验需要如此多的计划，所以申办方可能缺乏充分计划该设计的时间或资源。适应性临床试验在肿瘤科或罕见疾病中最常见。

一种解决方案是运用具有实时研究设计功能的随机化和试验供应管理（RTSM）技术，用户在几分钟内便可实施研究设计的更改。申办方和 CRO 可添加、禁用和更改治疗组合，基于新的治疗组合创建新的剂量规则并更改当前患者的治疗方案。

数据在适应性临床试验中的作用

一项试验中，最可能快速变化的两个因素是执行随机化的方式和选择对照组的方式。适应性临床试验是这两种变化因素的一个实例：在此类试验中，基于生物标志物将乳腺癌患者随机分至研究的一个特定组，但当获得更多的信息时，他们可能从该组转移至另一组。由于大数据，另有一个共享安慰剂组经过一段时间以后可与其他研究交叉引用。在适应性研究设计中，可以更快地完成 I-II 期试验，并缩短 III 期试验。结果是减少了成本和时间，并且暴露于实验中的患者更少。

Medidata OPAL Diagnostics 通过对行业数据的交互式基准测试来使用数据挖掘

数据挖掘概念适用于监管和非监管内容

临床试验不只是生成数据，还生成一系列**监管和非监管内容**。云计算内容平台可实现实时协作，从而更快地访问存档内容并以更简单的方式搜索和发现相关内容。

有何优势？

在临床试验中有来自试验主文件、合同、CV、IRB 信件、SOP 工作流等多种来源的多种内容。另有不断增加的内容，多名用户通过在线内容平台可实时地同时创建、编辑和批准多项内容。包含数据挖掘的各种基于云计算的技术平台有助于这些用户创建一个可随处访问且简便易用的交互式网络，该网络不仅可提高生产率，还可提高监管合规性和决策效率。

有哪些问题？

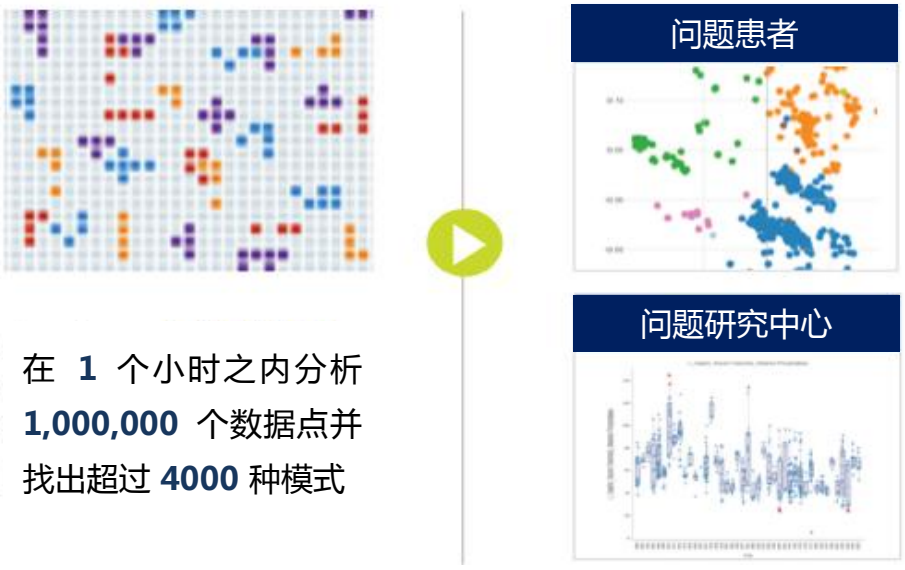
内容往往分散在各个独立来源中，这会使各组织管理和监管内容的方式发生改变。这在监管检查过程可能较为困难，因为必须能够迅速且准确地交付稽查员所需的内容。监管内容及元数据符合工作流程的 21 CFR 第 11 部分至关重要。

由于申办方和 CRO 需要多个平台管理监管和非监管内容，因此实施和维护这些系统成本很高。如果系统缺乏直观性，还会导致用户操作失败，并影响流程的合规性，从而可能导致监管机构合规性受到严重影响。您将如何管理众多不同的数据和信息来源，从而确保协作、监察准备和合规性？

监管内容管理的解决方案

拥有适用于**监管内容管理（RCM）**的集成端到端系统，提供所需流程并确保其机密性、高度完整性、可追溯性和可用性，是成功的关键。Zosano Pharmaceuticals 监管事务和质量部副总 Hayley Lewis 讨论了在临床试验领域采用高质量 RCM 解决方案的重要性：

“在这个行业里，几乎所有事情均受到监管，因此拥有可追溯性和适合的操作流程极其重要。将您的文档置于便于编写、审查和批准的安全、经过验证的环境中；这也使您能够将注意力集中在试验中的其他重要的关键领域。”



Medidata 中央统计分析 (Medidata Centralized Statistical Analysis) 采用机器学习方式推动基于风险的监察管理

在临床操作中利用数据科学

在研究期间采集临床患者数据的同时，还要获得具有价值的相关元数据：**治疗领域、研究中心位置、支付状态和各种流程持续时间**，当汇总数百或数千项其他研究的相关数据后，可获得用于分析的强大数据库。

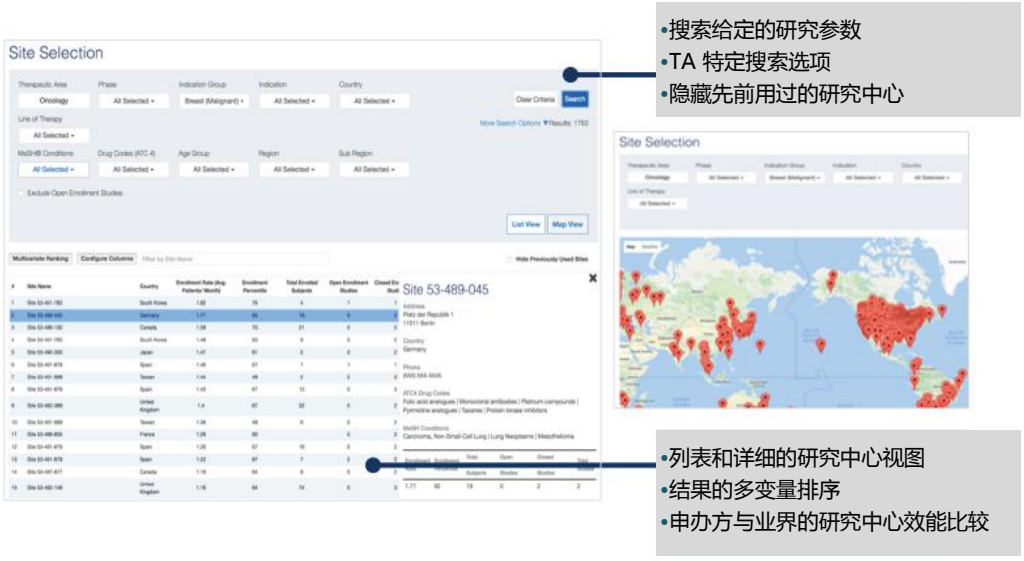
元数据有何优势？

临床操作中的预测分析有助于优化试验中的一些最重要的流程和基准测试程序。使用以往的研究中心绩效选择未来临床试验的研究中心可大大提高患者招募率并提前研究开始时间。采用机器学习异常检测可自动标记需要进一步调查的数据，从而实现基于风险的监察。在临床操作中利用这些类型的预测分析可缩短试验时间表、提高数据采集准确度并最终提高效率。

有哪些问题？

首先，采用预测分析成功采集元数据需要一个相当庞大的数据库和以往类似试验的大型资料库。此外，鉴于患者的隐私和保护，必须对此数据库进行标准化和匿名化处理。另需数据科学专业知识进行适当的统计建模，从而获得行动所需的关键见解。

利用能够实现临床试验元数据采集、汇总和分析的软件和解决方案，临床试验技术人员正在解决这些问题。使用具有基于风险的监察解决方案，申办方可全面扫描临床试验数据库中各个数据域、研究中心和患者之间的不一致。检测到异常后，研究中心监查员可直接定向这些异常，而无需进行高成本且耗时的 100% 原始数据核对。



Medidata OPAL 研究中心选择方案 (Medidata OPAL Site Selection) 使用研究中心绩效历史记录的基准分析程序帮助选择研究中心

结论

更多数据点的创建和可用性正改变着一切（从工作和休闲活动到身心健康）。如今，临床试验涵盖从传统临床数据到高分辨率图像、实验室检测、监管文件、基因组和移动医疗数据等所有方面。真正了解这三大技术（大数据、数据挖掘和数据科学）带来的效益和挑战，能够让我们将试验带入生活，从适应性临床试验、临床操作和监管内容管理方面获得更深刻、更有意义的见解和实际价值。这些是进行更加有效的试验的关键。

关于 Medidata

通过为临床研究提供业界领先的基于云计算的解决方案，Medidata 正在重塑全球药物和医疗设备研发的概念。通过先进的应用程序和智能数据分析，Medidata 帮助全球生命科学客户（包括近 850 家全球制药公司、生物技术企业、诊断和设备公司、领先的学术医学中心以及合同研究组织）推进科学目标。

Medidata Clinical Cloud® 为临床试验带来了全新的高质量和高效率水准，使我们的客户能够尽早和更快做出更明智的决策。我们拥有无可比拟的临床试验数据资源，由此获得的深刻见解可为未来的发展铺平道路。Medidata Clinical Cloud 是世界前 25 家全球性制药公司中的 17 家以及前 20 家全球性医疗设备研发机构中的 16 家都采用的主要临床试验技术解决方案（从研究设计、计划直至执行、管理和报告）。

info@mdsol.com | mdsol.com | +1 866 515 6044