NESTED KNOWLEDGE:

A MEDICAL RESEARCH PLATFORM FOR THE DIGITAL AGE

Whitepaper: From the Nested Knowledge Executive Team

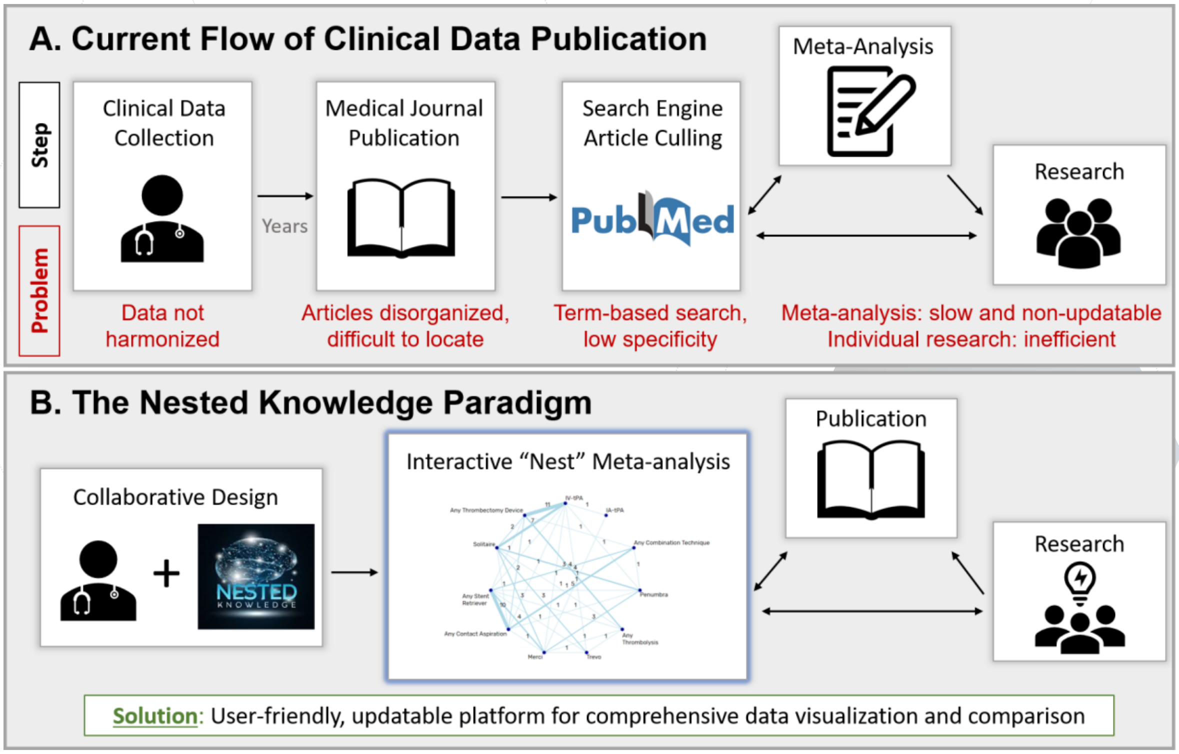
October 25th, 2019

Executive Summary

Evidence of therapeutic safety and efficacy is central to medical decision-making, but the outcomes of clinical research are communicated using opaque, outdated methods. Present-day medical literature — the foundation for clinical decision-making, physician guidelines, device/drug development, insurance reimbursement, regulatory approval, and health technology assessment across the globe — is presented in static PDF reports, whose data are not systematically organized or made accessible to users . No central data repository exists for a comprehensive evaluation of all clinical outcomes related to any given therapy or disease, either for patients, researchers, or medical professionals.

Nested Knowledge is a medical publishing and analytics platform that presents scientific research from thousands of studies in a dynamic, updatable, interactive visual interface. Our data-gathering augmentation technologies, including machine learning-driven data extraction, enable us to create structured, updatable datasets of clinical outcomes that we analyze and present using best practices in network meta-analysis. We are currently scaling our centralized database and data-visualizations to cover all published clinical outcomes research, thus becoming the system of record and central access point for all clinical publication. We are partnering with leading research organizations to apply our novel data-gathering methods across the diverse outputs that depend on medical research. If successful, Nested Knowledge will provide the only resource through which users throughout the healthcare value chain can make decisions based on all relevant published evidence.

OUR VISION is to gather all clinical outcomes data in a single, central system of record and enable treatment, regulatory, insurance, investment, and R&D decisions to be driven by comprehensive, evidence-based analysis.


A. Current Flow of Clinical Data Publication
Step
Problem
Clinical Data Collection
Years
Medical Journal Publication
Search Engine Article Culling
PubMed
Meta-Analysis
Research
Data not harmonized
Articles disorganized, difficult to locate
Term-based search, low specificity
Meta-analysis: slow and non-updatable
Individual research: inefficient
B. The Nested Knowledge Paradigm
Collaborative Design
+
NESTED KNOWLEDGE
Interactive "Nest" Meta-analysis
IV-tPA
IA-tPA
Any Thrombectomy Device
Solitaire
Any Combination Technique
Any Stent Retriever
Penumbra
Any Contact Aspiration
Any Thrombolysis
Merci
Trevo
Publication
Research
Solution: User-friendly, updatable platform for comprehensive data visualization and comparison

PROBLEM STATEMENT:

Medical technologies are advancing rapidly, and clinical trials and studies are the most important sources for communicating about the therapeutic efficacy of these novel tools. These technologies often deploy data reporting and analytics to benefit patients directly, demonstrating how beneficial data management can be to human health. However, unlike data-driven medical technologies, the medical literature—the primary venue for analyzing these technologies—has not adopted best practices in combining or promulgating data. While articles are now available online, the gold standard in medical research, systematic reviews of high-quality trials, are published in static, non-updatable, and heterogenous formats, and then scattered across thousands of publications.

To keep up with the information overload and track the newest and best therapies, clinicians increasingly rely on expert aggregation services, which bridge the gap in effective data presentation but depend on the opinions of leading researchers without any improvement in access to the underlying data. In a scientific discipline, and one that underlies trillions of dollars in healthcare expenditure and hundreds of millions of medical decisions every year, these vital data should not be communicated in anything less than a scientifically validated, transparent, analytics platform. As conventional meta-analyses/systematic reviews are unfit to communicate up-to-date, comprehensive, and useful information about drug/device performance, there remains a pressing need for a service that provides the vast range of parties interested in medical research with the vital data necessary to make informed decisions.

The key problems are:

- Publications utilize outdated PDFs with static text and tables to communicate complex medical outcomes; the reader effectively has no access to the underlying data from either primary studies or systematic reviews.
- Research is never actually up to date: all major fields see monthly additions to the literature but the average topic takes well over a year to see a meta-analysis published.
- Even systematic reviews focus narrowly on a single topic due to inherent space limitations, preventing researchers from providing nuanced answers for complex clinical scenarios.
- With no platform to compare studies directly, each reader must painstakingly search the entire medical literature — scattered across multiple journal publications and timelines — to answer their own specific questions and generally find only a subset of relevant research.
- Published data are the source of interest for regulatory submissions in all major medical markets, most reimbursement decisions, and many investment and R&D decisions across drug and device companies and their shareholders. However, each organization undertaking these decisions must search and gather the same data using the same work-intensive methods.
- Technologies ranging from simple data management solutions to search algorithms to machine learning have never been applied comprehensively to aggregating published study data.
- No organization has undertaken to become the system of record (the Bloomberg or Crunchbase) of published clinical outcomes research, leaving the field without trustworthy sources for comprehensive analysis.

THE OPPORTUNITY:

We are not alone in recognizing the gap in provision of comprehensive clinical outcomes data using modern software methods. The NIH, MDIC, AAAS, FDA, and National Academies of Science, Engineering, and Medicine have all recognized the need for improved data management and presentation in medicine. Along with these organizations, journal subscribers are actively demanding that medical publication be transformed from its current static, scattered form to dynamic, combined, instant, and direct access to medical data.

It is difficult to overstate the impact of improving the way we collect and analyze clinical data. These data are the primary source for many steps in the healthcare delivery value chain, from therapeutic development and approval to clinical and insurer decision-making. In a discipline with a central mission of advancing patient care through evidence, systematic record-keeping of patients' outcomes is the foundation upon which medical consensus can be established and novel hypotheses can be tested.

Data Access Methods:	**Current**	**Future**
Publication Style	Fixed	Dynamic
Combinations of Studies	Atomistic	Compound
Publication Process	Single-event	Continuous
Timeline of Availability	Delayed	Instant
Publication Medium	Journal article	Linked data & analysis

Scientific knowledge advances by testing hypotheses through replicated experiments. Currently, **replicability, transparency, and the scientific method** are practiced at the individual-study level, but scientific examination in the context of clinical medicine requires comparison across studies using similarly rigorous methods. However, the current approach to meta-analysis is too scattered and inconsistent to gather and analyze evidence using best scientific practices, crippling our ability to make conclusions or comparisons across studies. By combining and presenting data across medical disciplines with a focus on replicability and comprehensiveness, we propose that a scientific system of record—a structured platform for organizing all past data on any given clinical research question—would enable scientific principles to be applied not only to individual studies but to the entire process of planning, publishing, combining, and examining clinical evidence. This would provide a single platform for **experts to discuss the evidence** (without spending most of the time finding and comparing sources), for interested parties to **learn about the scientific consensus** (by finding all relevant data presented in a user-friendly interface with accompanying interpretations), and for **clinical science to be planned around well-communicated, structured evidence** examining relevant research questions and gaps in knowledge (rather than depending on each clinician scientist's scope of understanding).

In summary, the wider medical community has consistently called for clinical data to be combined and analyzed using scientific principles and novel technologies for gathering, analysis, and presentation.

THE NESTED KNOWLEDGE SOLUTION

Enter Nested Knowledge. We are a data analytics company focused on gathering all data from across medical disciplines and presenting it in a single meta-analytical platform. We have worked with leading clinical researchers to design a process that combines automation of best scientific practices with data-management support software to maximize the replicability and efficiency of data gatherers. The output of our data gathering is saved in a structured, updatable database, on top of which we have built automated statistics, data visualizations, and research tools that enable users to learn the broad conclusions of a medical field, and then also to drill down on any questions of interest through user-designed interfaces, or even access and download our raw data to complete independent analyses. We combine this scientific platform/database with comment tools that allow experts to interpret and discuss the data within the interface, building scientific consensus not just based on data-driven analysis but actually overlaying it directly on the evidence within our platform.

We have partnered with leading research institutions to design and validate our systems, which we publish systematically alongside our 100+ clinical coinvestigators. We continue to work with these institutions to improve our user experience, create new features, and refine our automation of best research methods. In February 2019, we completed our first proof-of-concept by comprehensively reviewing 350+ studies of acute ischemic stroke, the leading cause of preventable disability in the US. Since then, we have generalized and scaled our technology and database to cover cardiovascular and oncological disease states, and have expanded our research outputs to enable regulatory and insurance-focused reports to be generated directly from our interface. Because the breadth and detail of the data we have gathered, we have also had the opportunity to create novel data-quality and bias metrics driven by analysis of each study's data in context of all related studies and practices, and we are now focusing on building these metrics, as well as complete methodological documentation, directly into our interface so that users can scrutinize every step of the scientific process within our platform.

Our key area of technological advancement is the automation of the data collection process, which is not only quite costly in terms of manpower and time but is also difficult or even impossible to keep replicable without a single, centralized, software-controlled methodology. Even at leading institutions, an experienced data gathering team is generally expected to have error rates (missing or incorrect data) on up to 20% of data elements; by creating systems that induce regularity in data gathering practices and quality control checks, we have already reduced error rates to well under 10% while reducing time-per-study by over 40%. We are currently implementing a machine-learning method of direct data extraction from PDFs, and we anticipate that we can automate or augment over 80% of the data gathering process by the end of 2020, and will test our platform side-by-side against manual methods to demonstrate its accuracy and cost savings. These crucial technologies will allow us to scale rapidly and render the task of gathering millions of studies in one interface not only possible but entirely feasible.

The novel system of record for medical data that we are building will be a living systematic review that provides **1)** instant data access, **2)** structure and organization of study outcomes, **3)** data-visualization of everything from patient background characteristics to risk-of-bias analyses, **4)** searchability not just by MeSH tags or terms of interest but *by the actual data contained in candidate studies*, and **5)** a forum for discussion that enables interpretations to be connected to data of interest.

Our customers:

- **Providers, physicians, and patients:** When guiding patients through clinical decisions, physicians have insufficient tools for combining complex and nuanced information and presenting it in a form that can effectively inform patients.
- **Device and drug companies:** The impact, and therefore the value, of novel technologies is driven by evidence-driven clinical performance. Therefore, access to updated and comprehensive information to inform R&D and executive decision-making can help drive the development of life-saving therapies.
- **Reimbursement report generators:** Reimbursement, both public and private, for common therapies is driven by Health Technology Assessment Reports, which are based directly on the medical literature. More comprehensive and transparent data, gathered through scalable methods, would enable more informed and data-driven insurer reimbursement decisions
- **Regulatory writers:** Regulatory submissions, especially to the EU, include comprehensive literature reviews to supplement internal/preclinical data, and could similarly benefit from scalable, comprehensive data gathering technologies.
- **Biotech investors:** Biotech investments involve novel and often complex technologies, making clinical and commercial success difficult to predict. Using comprehensive clinical analytics to de-risk investment opportunities and survey the existing clinical tools and their effectiveness provides vital data-driven, independent feedback in advance of major investment decisions.

THE CHALLENGE AHEAD:

The principal challenge facing our centralized system of record is the sheer volume of heterogenous medical data currently published. Leading medical indices have over 30 million published articles indexed—each. Designing and scaling systems that enable nuanced, comprehensive comparison of therapies depends on identifying the subset of relevant studies with extremely high accuracy, followed by a seamless transition to data gathering, quality control, and output. We have deliberately chosen clinical studies as our central focus because of their greater impact on patient care, but also because consistent MeSH tagging, data presentation methods (most notably structured tables), and data division into patient characteristics, therapeutic methods, and outcomes enables a universal data schema to be applied across disease states. Based on the number of primary articles focused on humans (case reports, editorials, etc., as well as clinical studies, which comes to just over 560,000 per year) and meta-analyses/systematic reviews published per year (~22,000), there are roughly 8 million articles in the "universe" we are attempting to address.

Using current manual methods, a database of hand-gathered data from 8 million articles is effectively impossible. Even using expert researchers, this would take over 2,000 person-years to gather, and scaling such an effort would depend on complex coordination of research efforts across fields. Therefore, **the challenge Nested Knowledge has focused on addressing is technological**: automating or augmenting each modular task involved in meta-analyses. We invest over 70% of our time not in gathering studies but in carefully recreating the work of librarians, meta-analytical statisticians, and data gatherers in software. Thus, we hope to turn this impossibility into a reality, by reducing human effort to only the most vital interpretive aspects of medical research.

Our core competencies:

- Data gathering automation. We focus on: literature search optimization, PDF data extraction, database schema design, auto-updating, and automated management of research processes.
- Data visualization. Visual communication of evidence is much clearer than textual, and "nesting" visuals enables both high-level and detail-driven examination in a single platform.
- Transparency. Our automation also enables an unprecedented level of detail and consistency in reporting our methods from data gathering through to publication on our site.
- Statistical analysis. We use input from experts in health informatics, "big data" analytics, and meta-analytical statistics to create tools for completing meta-analytics through our interface.
- Medical expertise and partnerships. We ensure that subject matter experts are involved in the research and output design process for every meta-analysis, as these research questions can have nuances and field-specific needs that must be built into our systems.
- User-designed interfaces and feedback. Our initial visualizations were developed in partnership with leading researchers and potential users, and we now use in-site feedback mechanisms to get further data on how users interact with evidence.
- Updatability and Durability. Our systems are built to last, and to be fully updatable as new studies and even new fields of study emerge.

HOW WE GATHER AND PRESENT OUR DATA:

Our systems for data gathering represent a simplification, streamlining, and centralization of the many tasks needed in building a literature review. We start from study protocols, which are generated with librarians and subject matter experts, and input the search terms and inclusion criteria from this protocol into our system, which pulls search and study metadata and then automatically excludes studies based on structured criteria, citation networks, and simple text analysis. A researcher then completes exclusion of "fringe studies," and completes manual data gathering from all included studies. Simultaneously, our machine learning systems extract data from PDFs and quality-control the researcher, while also using each dataset gathered as training data to improve automated data gathering. Then, summary and inferential statistics can be generated within the interface, and the data can then be exported as: 1) data visualizations (see below), 2) regulatory-compliant literature reviews, or 3) any Word- or CSV-based format, based on user-provided templates.

Step 1: Search
- Our system scrapes medical indices and automatically applies inclusion criteria

Step 2: Data Gathering
- Researchers gather data into structured databases
- **ML tabular extraction**

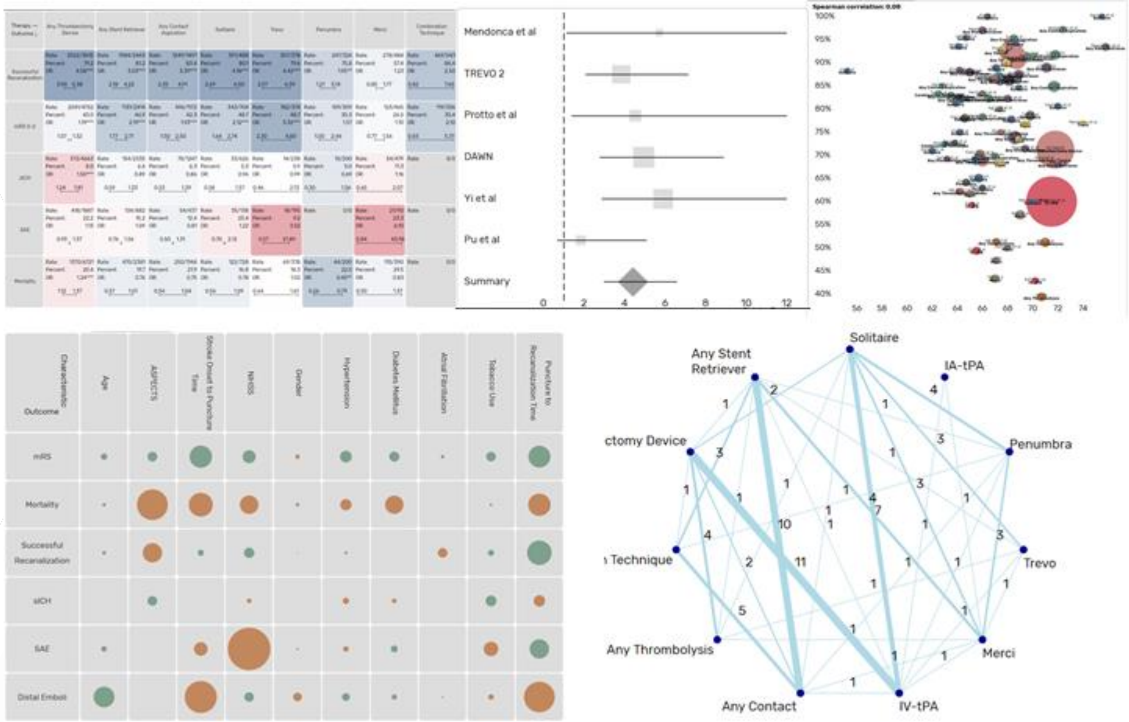
Step 3: Generate Output
- Automated summary & inferential statistics
- Formatted for **data-viz** or regulatory/insurance reports

Step 4: Auto-updates
- Our system re-runs the search daily
- Data gathering automatically initiated for new studies

Interactive Data-viz:

Nested Knowledge is available online at https://nested-knowledge.com/. We have the following features:

1. **Central "hub":** A data-heavy but simplified hub to access all visuals, with therapies compared based on odds ratios relative to best medical therapy
2. **Network diagram:** A network-meta-analysis inspired interactive diagram displaying the studies comparing different therapies with links directly to research tables
3. **Research tables** that allow users to view **all of the data** underlying our visualizations in its raw form, and select studies based on a range of criteria.
4. **Correlation matrices** that enable users to compare the impact of different patient characteristics on patient outcome, which link to **Causation plots** that display individual studies in an x-y graph for any given research question, with links to the underlying studies.
5. **Forest plots** with all studies reporting data for each cell pop up on scrollover from the main page, enabling users to see tailored plots for any research question of interest.



These presentations are supported by tools to **improve navigability/granular control**, and are soon to be supplemented with our risk-of-bias and data-quality visualizations. We are also adding risk-of-bias visualizations, as well as comment features for general users. Our main page allows users to search for studies, not through general search terms, but by selecting *the actual data that studies must contain to be included*, which our database then presents in our visuals.

We believe that **interactive data visualization** is the future of meta-analytical communication. The range of visuals enable higher-level examinations of device comparisons to be supplemented with patient characteristics, as well as drilling down on the actual underlying data. Better data density, clearer communication, and user-focused design make this flexible format a superior method of communicating vital data about key clinical outcomes and medical decisions.

WHAT ARE WE DOING NEXT? WITH WHOM?

Nested Knowledge is laser-focused on expanding our current literature review database, which is growing at a rate of hundreds of studies a week—and accelerating its growth by over 20% per month. We are also publishing scientific articles detailing our medical research methods and, more importantly, our machine learning-driven data extraction technologies alongside medical researchers. We are expanding first into cardiovascular therapies, followed by oncological therapies, based on which we will have coverage of over 50% of the causes for morbidity and mortality in the US. We are also partnering with a leading **health technology assessment report** organization, which it creates to support both insurance decision-making and society guidelines.

On top of scientifically examining our own internal processes, we are also focusing on user experience on our site. We are adding expert interpretation features that allow our coinvestigators to explain basic concepts and give insights on the data for general users, and will follow this with general user-comment and comment-rating features so that our platform can provide not only all relevant data but also a forum to discuss the evidence. Lastly, we are seeking a broad range of clinician coinvestigators interested in coauthoring and guiding literature reviews that will endure into the future as the central source related to each subject field of interest. We look forward to updating the medical public on our progress toward a central system of record for medical data, and enabling clearer, more scientific examination of the whole of clinical research to inform medical decision-making and, ultimately, save lives.