***Attendees:***

* **BMGF:** Alison Jones, Steven Kern, Thea Norman, Tom Peppard (*also CC’d: Laura Lamberti, Anita Zaidi*)
* **EuPathDB:** Brian Brunk, Omar Harb, Jessica Kissinger, David Roos, Chris Stoeckert
* **GEMS:** Anna Roose (*also CC’d: Karen Kotloff, Mike Levine, Dilruba Nasrin, Sharon Tennant*)
* **MAL-ED:** Michael Gottlieb, Dennis Lang (*via teleconference*), Jessica Seidman, David Spiro, Karen Tountas (*also CC’d: Danny Carreon, Viyada Doan, Eric Houpt, Stacey Knobler, Monica McGrath*)

***Agreements:***

1. Agreement that database resources supported by this award should be constructed so as to readily accommodate datasets beyond those that are EDD-specific. Organizing in this way should make it possible for a user to query across available datasets (EDD, malaria etc), when appropriate.
2. Agreement that “release 0” encompass GEMS and MAL-ED data already organized/structured in HBGDki.
3. Agreement that first release of data (“release 0”) to this grant’s internal team will take place by August 2, 2017.
4. Agreement that a small group be convened in August to look at the “release 0” prototype and to draft policies and procedures for data access, and further plans for data loading, analysis, etc.
5. Agreement that we should target Nov-2017 for public presentation and/or release of “release 0”, perhaps in association with the ASTMH meeting in Baltimore (or elsewhere, if more appropriate)
6. Agreement to the schedule below
   1. August 2017 (or earlier, if possible): internal release ("release 0") for beta testing
   2. August 2017: face-to-face meeting of core team to start using internal release and to draft policies and procedures (location TBD).
   3. Nov 2017: Penn team demo and/or public release of EuPathDB ("release 0") at ASTMH (with possibility of opening site for use during the meeting?).
   4. Oct-Dec 2017: investigator workshop (location TBD). BMGF to lead on planning logistics.

***Action Items:***

1. Action Item: David/Alison to send email and/or schedule a call so that grants administrators at Penn can have a few questions answered *… done*
2. Action Item: David to provide updated results framework and tracker (by April 30)
3. Action Item: Tom to schedule a call with Thea, David, and others from QS and EuPathDB as appropriate, focused on understanding HBGDki datasets and tools to help guide this partnership and figure out what from HBGDki could be leveraged to mutual benefit.
4. Action Item: Thea/Steve to discuss how MAL-ED grant completion affects the EuPathDB project and what if any actions are required (for example to ensure effective engagement in DB design and testing as well as providing context and definitions to the datatype variables).
5. Action Item: David to create a cloud-based doc to collect input from attendees post-meeting for likely use cases that the EuPathDB needs to anticipate. Attendees to categorize their questions: just for one study, for both studies, most compelling questions. *Link:* <http://tinyurl.com/ClinEpiDBusecases4EDD>
6. Action Item: Tom to provide Penn team with HBGDki information on GEMS data *… done*
7. Action Item: Anna and Michael/Karen to contact respective investigators to get a short list of those who (1) are deeply familiar with data set and (2) would be willing to serve as part of a beta testing team.
8. Action Item: Thea, Laura with guidance from Steve, Anita, Penn and GEMS/MAL-ED leads plan an August meeting that includes a core set of GEMS/MAL-ED investigators or local researchers.
9. Action Item: Thea to publish notes and action items from the meeting *… this document*

***Discussion & Notes:***

**Roles, Strategy and Background**

* As of Jan-2017, all BMGF recipients must publish their results in an open access journal.
* BMGF Global Health Data Access has the following benefits: innovation, collaboration, efficiency, accountability, capacity strengthening.
* For Global Health Data Access to be effective, BMGF is working to create data-hosting hubs so that researchers can easily find relevant data, and so that similar study data sets are integrated and can be interrogated in parallel.
* This purpose of this grant is to create an EDD hub for BMGF-funded data.
* Both MAL-ED and GEMS had the requirement in their original grants to make their data open at some (previously unspecified) date.
* This project is intended to satisfy requirements of the MAL-ED and GEMS grants to make data public.
* Metrics of success from the BMGF perspective
  + Use of the database by BMGF staff, EDD project researchers / data providers, and the global research community (including endemic country researchers) to access GEMS & MAL-ED datasets, advancing EDD research
  + Use of these datasets to identify gaps in available information, and design studies aimed at filling those gaps.
  + For BMGF to obtain insights from the outcomes of this grant that further inform the Foundation’s approach to making data publicly available and useable.
* Datasharing: the HBGDki effort currently hosts data covering 10M children. Collecting these data was a major effort, and data is currently available only to participants in the HBGDki (this was necessary to encourage investigators to contribute data). Planning is underway to refresh data use agreements so as to allow for more open access where possible.
* This project creates an opportunity for some HBGDki data sets to also be hosted in a public repository.
* Tom Peppard is the connection to the HBGDki Initiative and tools from HBGDki that are available to this partnership

**Review of Grant Agreement (Alison Jones)**

* Alison Jones (Alison.jones@gatesfoundation.org) is the point of contact for all questions on the grant
* Kathleen Parsons (Kathleen.parsons@gatesfoundation.org) is contact for meeting scheduling.
* Any increase/decrease to a budget line of more than 10% must be approved by Thea in advance in writing.
* Action Item: David/Alison to send email and/or schedule a call so Penn grants administrators can have a few questions answered *… done*
* BMGF is open to a public press release announcing the grant, but advance approval is required, especially for any use of official BMGF logo.
* Social media also requires approval.
* David looking for input from meeting attendees on timing for near-term deliverables as well as how to prioritize what goes into the iterative database releases that are anticipated.
* Alison clarified that the results framework and tracker is a living document that should absolutely be updated to fit the scientific objectives, and that there is no need to maintain detailed backwards linkages to original proposal.
* Timeline for deliverables on this project helps BMGF manage this investment within the envelope of its internal budgets.
* Action Item: David to provide updated results framework and tracker (by April 30)

**GEMS Project Overview (Anna Roose)**

* GEMS (2007-2011) was conducted in seven high-mortality developing countries: 4 African and 3 Asian. GEMS-1a (2011-2013) conducted in same countries
* Study Design
  + GEMS enrolled cases with moderate to severe diarrhea (MSD): 8-9 per age stratum per fortnight enrolled at Sentinel Health. Three age strata: 0-11mo, 12-23 mo, 24-59 mo
  + Controls without diarrhea enrolled in same community
  + Both cases & controls were followed 60-days later for mortality.
  + Separately, serial healthcare utilization and attitudes surveys (HUAS) were conducted in the same communities to estimate the proportion of children in the community who experienced MSD within the past week; if so, determine whether they sought care at a SHC within 7 days of diarrheal onset.
  + R value is shockingly low: less than 33% end up seeking medical care for MSD
  + GEMS-1A was similar to GEMS, but cases were less sever (mild/moderate diarrhea).
* Datasets
  + High dimensional data sets. The complexity of the GEMS data sets makes them difficult for new investigators to learn how to use the data.
* To ensure comparability across sites, there was training that took place at all the sites and provided ELISA kits from a single supplier. Most of the testing was just done at the sites and was done in standardized way.
* There are certain derived variables in addition to underlying data. Having these variables available upfront will be helpful for new investigators to use. At EuPathDB, they make derived variables available: these can serve as useful tool for QA.
* VIDA is a follow-on study to GEMS in Mali, the Gambia and Kenya. Focused on understanding the impact of the rotavirus vaccine on MSD.
* David looking for input from GEMS team about which other data sets that they are aware of would be complementary and valuable to host in EuPathDB, including but not limited to GEMS follow-on studies. Note that early discussion is helpful, so as to help in study design and plan for what to expect, and when.
* HBGDki has >100 datasets focused on stunting. Spanning genetic, anthropometry, clinical, observational.
* Action Item: Tom to schedule a call with Thea, David, and others from QS and EuPathDB as appropriate, focused on understanding HBGDki datasets and tools to help guide this partnership and figure out what from HBGDki could be leveraged to mutual benefit.

**Mal-ED Project (Michael Gottlieb & Jessica Seidman)**

* FNIH received first MAL-ED BMGF funding in late 2008. FNIH collaborated with BMGF to build out data access plan.
* Observational community-based study, 8 sites (South America, Africa, Asia), 200 children (age 0-2years and subsequently up to 5 years per site). Kids enrolled within first 17 days of birth and followed for 4 years.
* 2,145 children enrolled overall.
* Outcome measures: growth, cognitive development, vaccine response, biomarkers of gut function.
* Bulk of data is 2x weekly household visits: illness surveillance and diarrhea stool sample collec­tion. Most were mild cases of diarrhea that do not last long.
* Stool testing: enteropathogen detection including PCR and enteropathic / fecal inflammatory biomarkers. Testing done on quarterly samples; other monthly samples archived.
* UVA is now retesting stool samples using TAQ. This will be an important data supplement to pull into the database eventually. Anna shared that TAQ data also exist for GEMS datasets.
* MAL-ED has very different participant characteristics than the GEMS data which was focused on moderate to severe diarrhea.
* No interventions: strictly observational.
* Extensive “recipe” database is a result of this study. These are detailed food intake questionnaires that could be scored to estimate macro-/micro-nutrient intake quantities. The questions & responses are of high resolution. For example there are many (100s or possibly 1000s?) of “recipes” for tea that were collected from mothers.
* Blood collections include measures for assessing immune responses to recommended childhood vaccines; micronutrient assays and lead levels.
* Urine lactulose/mannitol (L/M) assay and urinary iodine
* Note that at months 6 & 15 urine was collected as a 5-hr sample (to support L/M testing), and at 24-months it was a spot sample at some sites since L/M assay was optional at that visit. Jessica mentioned this as an example of study procedure details that might not be obvious from the data that will be important to capture in the database.
* There is understandable sensitivity regarding MAL-ED cognitive data, which is therefore not expected to be hosted on EuPathDB without further discussion. Prior to release on EuPathDB, a plan needs to be in place that covers how access takes place and how to insure that users know how to use the data appropri­ately.
* Interactions with data providers will be critical.
* Raw and cleaned data are available from J Seidman, as well as at individual study sites. D Roos suggested that cleaned data should be the first focus for hosting on EuPathDB.
* Paper forms reside at sites,. Will assume that what the MAL-ED team provides constitutes the dataset (as opposed to creating a mechanism that goes back to the paper forms or to the field workers)
* BMMI companion-project (separately funded, Jeff Gordon PI): breast milk and stool samples being analyzed including a maternal stool sample collected just after birth. A lot of this data will be coming out. Should be considered for inclusion in this project. (BMMI subjects are a subset of the MAL-ED study thus provide all of the MAL-ED data elements in addition to those unique to the BMMI protocol.)
* Currently closing databases for completion of the study.
* MAL-ED funding all comes through BMGF grant: the grant wraps up over the next 18 months.
* Action Item: Thea/Steve to discuss how MAL-ED grant completion affects the EuPathDB project and what if any actions are required (for example to ensure effective engagement in DB design and testing).
* With a grant supplement from BMGF, enrolled children were further followed from 25-60 months of life primarily to assess cognitive and related developmental outcomes.

**Platform Demonstration (Brian Brunk)**

* Infrastructure improvements under development for other components of the EuPathDB Clinical & Field Metadata Integration Platform (e.g. for data loading and UI) will also benefit this project; similarly, effort invested in this project will benefit other related resources.
* Analysis tools can readily be hooked into the EuPathDB environment; for example, a Shiny app was demo’ed.
* Anticipate that GEMS and MAL-ED data will be loaded / hosted in a customized way to enable each dataset to be interrogated individually, so as to facilitate QA and ensure maximal use of each dataset. Overlaps will also be leveraged, however, to permit cross-dataset queries where appropriate. Visualization tools should permit users to browse these data, including areas where the two data sets intersect.

**Use Case Discussion (David Roos)**

* Will require careful attention to determine to what extent it is possible to query GEMS and MAL-ED in a single search.
* Action Item: David to create a cloud-based doc to collect input from attendees post-meeting for likely use cases that the EuPathDB needs to anticipate. Attendees to categorize their questions: just for one study, for both studies, most compelling questions. *Done … here is link:* <http://tinyurl.com/ClinEpiDBusecases4EDD>
* Draft Use Cases for interrogating EDD datasets *… use above link to review/ revise / add / share:*
  + Which children in the MAL-ED study had no diarrheal episodes whatsoever over the first 24 months of life?
  + What children (or what percentage of children) show no evidence of exposure to *Shigella* (or *Cryptosporidium,* or whatever) by age 5?
  + What proportion of children have never been infected with an enteric pathogen during the first two years of life?
  + What is the effect of delivery at a health care facility vs. at home on health outcomes?
  + What micronutrients are inversely correlated with EPEC infection, or with *Firmicutes* abundance in the gut microbiome?
  + How does the duration of a diarrheal episode correlate with the duration of the *previous* diarrheal episode in the same child? How is this affected by age, etiological agent, *etc*?
  + What is the likelihood of seeing multiple patients with severe diarrhea from the same household within the same week? How is this affected by geographic location, calendar month / season, *etc*?
  + What is the cumulative enteric pathogen burden up to 7 months of age in Dhaka (or Bamako, or Iquitos, *etc*)?
  + How does cumulative enteric pathogen burden impact response to polio vaccination?
  + What is the correlation between neutralizing antibody responses to rotavirus vaccination *vs* poliovirus vaccination?
  + What is the distribution of pathogens associated with different clinical phenotypes; for example: what fraction of watery diarrhea is associated with rotavirus (and at what titers, if quantitative information is available)? How do these numbers change with age, geographic location, diet, primary water source, recent rainfall, etc?
  + How common is it for children with low HAZ (height-age Z-score) to display a growth burst within the subsequent 6 months? Identify children in this category for which blood samples are available? Was the same phenotype observed for other children in the same household?
  + What is the correlation between exposure to distinct pathogens and cumulative growth (in the MAL-ED cohort)?
  + What is the correlation between bacterial beta diversity in microbiome samples and growth?
  + Is higher beta-diversity in the maternal breast milk microbiome associated with higher growth during the first year of life?
  + What is the prevalence of distinct pathogens the community (MAL-ED cohort) vs the clinic (GEMS)? For those that display a difference, is this consistent with moderate vs severe diarrhea in GEMS1 vs GEMS1a?
* Note that we do not currently envision providing rigorous support for statistical analysis or sample size calculations, but rather to simply provide clues that will require more detailed discussions with statisticians & epidemiologists to assess significance.

**Review and Discussion of Investment Outcomes**

* Will data obfuscation be required?
  + MAL-ED: all de-identified except for a single survey question. Child and country are listed as IDs. Date of birth is not considered as identifiable: date of birth will be useful in the context of seasonal considerations.
  + GEMS: what is being transferred to EuPathDB is all de-identified.

**Data and Data Integration (Jessica Kissinger)**

* The process has begun … significant transfer has already taken place for both GEMS and Mal-ED.
* Site will be private while under development: release will only happen once stakeholders sign off.
* Agreement that first release of data to internal team will take place by August 2, 2017.
* EuPathDB team will not dictate the timing of public release. Next steps would be for the team to decide what should go into the initial release and who should be included for the internal release.
* Additional datasets
  + Quantitative assessments of enteric pathogens in diarrhea & surveillance stools
  + MAL-ED companion studies (breast milk and microbiome)
  + Aflatoxin effects – Dr. Felicia Wu (PI), Michigan State University; Sani-Path – Dr. Christine Moe (PI), Emory Universiy.
* At present, EuPathDB team knows nothing about availability or location of physical samples, but can accommodate such information.
  + Need linking of archived bio-samples so that users could be aware of potential for future collaborative studies.
  + For MAL-ED, the sites own the samples. Michael Gottlieb agrees that a EuPathDB users should be able to find out if samples are available. This should be discussed with site investigators.
  + MAL-ED team in DC does not have details on sample availability.
  + Having data flagged to indicate that physical samples were collected is an important collaboration tool and also can help researchers understand what physical samples may be available for future collaborative studies.
* Do not envision tracking sample availability over time … more as a way of enabling a collaboration between database users and study investigators.
* HBGDki has a more structured form for the GEMS data.
* Action Item: Tom to provide Penn team with HBGDki information on GEMS data *… done*
* Agreement that “release 0” is what is already organized/structured in HBGDki for GEMS and MAL-ED. Plan the first public release around “release 0.” “Release 1” will extend on this. Once “release 0” is internally shared, group to evaluate if public release should take place sooner than Nov-2017.  
  Tom Peppard already sent Penn team the data dictionary that the GEMS team shared with the HBGDki.

**Development of formal ontologies relating to enteric disease (Chris Stoeckert)**

* The HBGDki data standard is based on CDISC. Tom can provide a structured description to help understand relationship between raw data and CDISC-formatted data.
* Still need to understand if it should be assumed that all CDISC-formatted datasets are integrate-able in a useful way.
* Formal ontologies will best support community expansion for this project. Note that CDISC is not a formal ontology; rather it is a data standard. It describes how to structure the data and includes some controlled terminology, but it lacks the detailed relationships among and descriptors for individual data fields.

**Outreach activities (Omar Harb)**

* EuPathDB has a team, led by Omar, focused on outreach for EuPathDB projects.
* Outreach will include training research in developing countries: focused on empowering local users and making sure they can be successful accessing EuPathDB using their own computers in their respective settings.
* Steve: strong support for robust outreach strategy to support the success of pubic hosting.
* Action Item: Anna and Michael/Karen to contact respective investigators to get a short list of those who (1) are deeply familiar with data set and (2) would be willing to serve as part of a beta testing team.

**Question of scope and vision for the project (All)**

* Can EDD database area be set up as part of a Clin/EpiDB housing other datasets (non BMGF, non EDD)? This way, a single URL would take users to a master list of datasets, allowing users to drill down into EDD-specific data, or other areas (e.g. malaria) … and also to extract EDD information from malaria studies (for example).
* Agreement that database resources supported by this award should be constructed so as to readily accommodate datasets beyond those that are EDD-specific. Organizing in this way should make it possible for a user to query across available datasets (EDD, malaria etc), when appropriate.

**Planning for MAL-ED and GEMS Investigator meeting (All)**

* Suggestion to target Nov-2017 ASTMH, but Karen notes that most MAL-ED investigators do NOT attend ASTMH. EuPathDB will be presenting at ASTMH, so there is still the opportunity to highlight this project: could be publicly available during the meeting or maintained as public thereafter.
* BMGF can provide a convening budget to make the investigator meeting happen in a way that optimizes participation from study investigators/epidemiologists/statisticians.
* Agreement that a small group be convened in August to look at the “release 0” prototype and to draft policies and procedures for data access, and further plans for data loading, analysis, etc.
* Action Item: Thea, Laura with guidance from Steve, Anita, Penn and GEMS/MAL-ED leads plan an August meeting that includes a core set of GEMS/MAL-ED investigators or local researchers.
* EuPathDB does not currently require all users to log on, unless they want to save their query history, receive updates, etc. That said, access to this resource can be set up differently. Michael G said that investigators have signed off on making the data publicly available: going forward he wants to be transparent with them regarding the details.
* GEMS has been over for a long time: investigators do not convene anymore. Nov-2017 would be great time to introduce VIDA investigators to the EuPathDB tool.
* Agreement to the schedule below
  + August 2017 (or earlier, if possible): internal release ("release 0") for beta testing
  + August 2017: face-to-face meeting of core team to start using internal release and to draft policies and procedures (location TBD).
  + Nov 2017: Penn team demo of EuPathDB ("release 0") at ASTMH (with possibility of opening site for use during the meeting?)
  + Oct-Dec 2017: investigator workshop (location TBD). BMGF to lead on planning logistics

**How We Expect To Work Together**

* David has requested meetings with MAL-ED and GEMS leads to support efforts between now and August 2017.
* MAL-ED DCC will need to be available to Penn team to help prep internal release.

**Wrap up and Next Steps:**

* Action Item: Thea to publish notes and action items from the meeting *… this document*